Contextual Features for Automatic Essay Scoring in Portuguese

Lucas Galhardi[1], Maria Fernanda Herculano[2], Luiz Rodrigues[2], Péricles Miranda[3], Hilário Oliveira[4], Thiago Cordeiro[2], Ig Ibert Bittencourt[2,5], Seiji Isotani[5], and Rafael Ferreira Mello[6(✉)]

[1] SENAI PR University Center, Londrina, Brazil
[2] Center for Excellence in Social Technologies, Federal University of Alagoas, Maceio, Brazil
[3] Federal Rural University of Pernambuco, Recife, Brazil
[4] Federal Institute of Espírito Santo, Vitoria, Brazil
[5] Harvard Graduate School of Education, Cambridge, USA
[6] Centro de Estudos Avançados de Recife, Recife, Brazil
rafael.mello@ufrpe.br

Abstract. Automated Essay Scoring (AES) efficacy often varies across linguistic and contextual nuances. This study addresses this gap by proposing and evaluating a contextualized approach tailored for Portuguese. Unlike prior research, which often focused on overall scores or limited to general-purpose features, we explored devising contextualized feature extractors and investigated their impact on predictive performance. Our analysis encompassed the proposed specific features (conjunctions, syntactic quantification, and entity recognition) and two well-established baselines (i.e., TF-IDF and Coh-Metrix). Utilizing the Essay-BR dataset (n = 6,563 essays), we investigated our approach through classification and regression tasks supported by diverse machine learning algorithms and optimization techniques. Mainly, we found that our approach enhanced predictive performance when combined with existing techniques. Our findings reveal the importance of addressing and considering contextual nuances in AES, revealing insights that might help accelerate the evaluation of essays in a large-scale setting.

Keywords: Essay Scoring · Natural Language Processing · Learning Analytics

1 Introduction

The National High School Examination (ENEM) is a key assessment in Brazil. It evaluates students' educational competence after completing the foundational education phase, playing a pivotal role in enabling many students to pursue higher education opportunities, including admission to public universities and access to financial support for private institutions[1]. Among its components,

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem.

A. M. Olney et al. (Eds.): AIED 2024 Workshops, CCIS 2150, pp. 270–282, 2024.
https://doi.org/10.1007/978-3-031-64315-6_23

ENEM features a discursive-argumentative essay, in which students must address a proposed problem, expressing their viewpoint and concluding by proposing an intervention to mitigate or resolve the stated problem. This essay must have up to 30 lines and its assessment is based on five competencies, such as *proposal intervention*, *organization*, and *coherence and cohesion* [7]. Each of these criteria concerns a competency, and scores are assigned on a scale ranging from 0 (indicating a complete lack of mastery of the mode) to 200 (indicating excellent mastery of the mode). Consequently, ENEM's scores range from 0 to 1,000 by summing up the scores for all five competencies.

Following ENEM's application, two evaluators are responsible for reviewing these essays. Evaluators are tasked to evaluate "*a perspective substantiated by coherent and well-structured arguments with logical consistency.*" In cases where there is a significant divergence in scores, a third evaluator is brought in to reevaluate the essay, to achieve a potential consensus on the final score[2]. However, the manual grading process - although indispensable - is well-recognized for its notable limitations related to the fatigue evaluators are likely experience due to its repetitive nature. Furthermore, because it relies on human judgment, the grading process is subject to various inconsistencies and biases, leading to an inherently unreliable assessment.

A potential solution to this challenge is known as Automated Essay Scoring (AES). In practice, AES might partially automate the essay evaluation to improve the efficiency of evaluators while guaranteeing a consistent, impartial and coherent grading outcome [13]. Most often, AES relies on Natural Language Processing (NLP) and Machine Learning (ML), where regression and classification models are the primary approaches [11]. Particular to ENEM's context, the emphasis on feature-based approaches is evident, demonstrating promising results on AES tasks given the full essay [10]. However, there is a lack of research addressing the assessment of the five competencies individually.

Given that the standard evaluation is split by component, understanding factors affecting each component's score will likely empower evaluators better. Therefore, this work focuses on ENEM's Competence 5 (C5), which is associated with the intervention proposal of the text. Thus, unlike prior research focused on the full essay and overall features, this study investigates the intervention proposal and its distinctive characteristics. For this, we propose three C5-specific features ("approach" hereafter) for the automated assessment of C5. The proposed features explore conjunctions and Named Entities Recognition (NER), informed by the National Institute of Educational Studies and Research (INEP), and analyze the syntactic structure within the essay's conclusion. Thus, we aim to extract informative features from the text to improve the learning model's performance in estimating the C5 score.

For this, we approach estimating the C5 score as both classification and regression tasks based on ML algorithms were applied to the extended Essay-BR dataset [10] (n = 6,563). The algorithms were assessed considering differ-

[2] https://www.gov.br/inep/pt-br/assuntos/noticias/enem/conheca-o-processo-de-correcao-das-redacoes.

ent combinations of the proposed and general-purpose features (e.g., TF-IDF e NILC) and the stacking approaches. The results demonstrated that the features extracted by the proposed approach, when combined with general-purpose ones, led to significant enhancements in the AES performance.

Therefore, this paper contributes towards mitigating disparities in AI progress across regions. Whereas research has shown substantial progress in AES for English text, there is a worrying lack of research on other languages [2]. As a continental-sized, Portuguese-speaking country, Brazil and similar countries demand research attention from the AIED community. Hence, given ENEM's relevance for the country, developing AI systems that help optimize essay scoring is paramount. Thus, this paper takes a step towards fulfilling that gap by introducing ENEM-specific features that, based on our experimental results, contribute to AES in the context of ENEM, specifically for C5.

2 Related Work

Research on ML for text evaluation is widely focused on the English language [2], creating a significant disparity in the use of AIED across regions, given that those whose language is not English are left unattended by AIED systems. Consequently, AES for Brazilian Portuguese, as addressed in the recent literature [9], represents a promising field, given the difficulty in efficiently and impartially evaluating a discursive text, as well as the limited research [2].

On the one hand, research related to the assessment of ENEM essays does not analyze scores on a competence basis, considering only the overall score [3]. The issue with this approach is that ENEM's score must be given for each of its five competencies[3]. Therefore, such a general assessment holds the limited potential to contribute to humans evaluators.

Some studies measure the performance per competence but employ the same features for all of them, lacking a specialized proposal [1]. The issue with this approach is that ENEM's evaluation involves contextual nuances, which are detailed for each of its components[4]. Thereby, developing AES systems limited to general-purpose metrics might achieve limited performance by not considering these nuanced insights.

For instance, Santos et al. [14] proposes an approach for Enem's fifth competence (C5). However, it uses the scores from the other four competencies to estimate the C5 grade without addressing its particular characteristics. Moreover, Fonseca et al. [5] personalize the assessment on a competence level but do so experimentally rather than proactively. It employed a feature selection procedure for each competency, retaining only features with a Pearson correlation of at least 0.1. More recently, some studies have directed their efforts towards specific competencies, such as textual cohesion [13], proposing specialized features that include characteristics from established tools like Coh-Metrix [4].

[3] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem.

[4] https://www.gov.br/inep/pt-br/assuntos/noticias/enem/conheca-o-processo-de-correcao-das-redacoes.

As the focus of this study, we concentrated on research that deals with (or at least mentions) C5. One study was found [12] that almost aligns with this aim: a systematic literature review of research identifying and automatically evaluating proposals for intervention in argumentative discursive texts. The review identifies five studies, one of which has been previously presented [1]. While related to the topic, the other four works do not assess C5 according to the scoring levels defined by ENEM; they solely perform binary identification of an intervention proposal. Furthermore, the techniques addressed by these articles focus exclusively on identifying textual markers that indicate the presence of arguments in the text.

With the introduction of the publicly available Essay-BR corpus [10], a new avenue for developing research emerges, mainly focusing on individual competencies. This prospect aligns with ENEM's current assessment framework. The work by [11] presents three assessment approaches: two generic ones based on textual representation (embeddings and recurrent neural networks) and one using feature engineering, with a distinct set for each competency. While achieving favorable results in C5 with the generic approaches, the feature engineering set for C5 exhibited the poorest performance compared to other competencies.

In summary, some studies automatically assess the total score of essays, some focus on other competencies of the ENEM, others relate to the proposal for intervention but fall outside the scope of the ENEM, and finally, studies that propose approaches for C5 but do not adhere to the current official criteria or did not achieve satisfactory performance in the task. Thus, this study aims to investigate C5 based on the official guidelines and evaluation criteria, integrating established and general approaches with aspects motivated by the intrinsic characteristics of intervention proposals. We contribute a contextualized approach towards helping improve AES systems for ENEM, a large-scale assessment with crucial implications for students, besides contributing to research on the Portuguese language, in contrast to prior research often focused on English speakers.

3 Proposal

To enhance the estimation of C5 scores, we conducted an in-depth analysis of its official guideline[5] (guideline hereafter), which offers a comprehensive description of C5's assessment process and serves as the reference for human assessors. Based on that, we propose and investigate multiple features for C5-specific feature extraction to improve the estimation of C5 scores, as detailed below.

The ENEM guideline for C5 describes that evaluating the intervention proposal focuses primarily on identifying the proposal in a general context and five additional elements. Mode/Form, Effect, and Elaboration are described next, whereas Action and Agent are detailed in the subsequent subsections.

In identifying the intervention proposal, emphasis is placed on the use of verbs such as "*dever*" (should) and "*propor*" (propose), with due consideration given to the identified and preserved tense (e.g., "*propõe*" (proposes) signifies

[5] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/outros-documentos.

a proactive action that has not been developed by the author of the text but rather cites another source proposing something). Note that there is a cautionary note against generic constructions such as "*é preciso*" (it is necessary) and "*é necessário*" (it is necessary), as they indicate that the author suggests the need for a proposal vaguely.

Hence, for the general intervention proposal, we devised two features to model the possible categories of terms: positive and negative. Such strategy considers that the identification of correct terms (*deve propor* - should propose, *proponha* - propose) signifies the presence of the element, whereas the presence of a null counterpart (*propõe* - proposes, *é preciso* - it is necessary) indicates the opposite. The features comprise a list of terms extracted directly from the examples in the assessment guideline.

In identifying the mode/form, there is no division into positive and negative features. Instead, a list of terms may include a mode/form. These terms encompass compound expressions, such as "*por meio de*" (by means of) and "*pelo qual*" (through which), and singular terms, such as "*através*" (through) and "*pelo*" (by). Note that singular terms were tokenized to prevent their inclusion within other words that might contain them.

Concerning "Effect," for which terms such as "*para que*" (so that), "*para isso*" (for this), "*objetivo*" (goal), and similar expressions were selected. Finally, for the "Elaboration" element, a feature was developed that encompasses six possible categories: exemplification, such as "*por exemplo*" (for example); explanation, such as "*na medida que*" (insofar as); justification, such as "*afinal*" (after all) and "*porque*" (because); contextualization, such as "*de forma a*" (in order to); addition, such as "*juntamente*" (together); and uncertainty, such as "*assim*" (thus). These categories were developed to capture various forms of elaboration within the text.

Conjunctions Quantification: Besides features directly based on the elements defined in the assessment guideline, we noted that they used *conjunctions* extensively. These words establish connections between other words and indicate the nature of their relationship. Conjunctions can be categorized into 15 groups, each dealing with a specific aspect (e.g., a cause-effect relationship). Then, extensive lists of Portuguese language terms for each of the 15 conjunction groups were selected and directly used in the extraction process from the essays. The features were extracted through simple counting and normalized by the original text's size (in terms of word count).

Overall Syntactic Quantification: We also investigated the text's syntactic organization based on syntactic quantifications established using the NLTK library[6]. In this analysis, only the essay's last paragraph, which was tokenized into individual words, was considered, thereby identifying the basic units of the text. Subsequently, the words were tagged with their respective part-of-speech (POS) tags to obtain information about the grammatical class of each word. Next, a frequency count of each tag in the paragraphs was conducted. This

[6] https://www.nltk.org/.

allowed identifying the most frequent grammatical classes in the text and a better understanding of its linguistic structure.

Named Entities Quantification: Recall that a sound intervention proposal includes action, medium, agent, and effect. Therefore, we developed a model that identifies and quantifies these elements in the text. For this, the spaCy library and the Prodigy framework were used. A new database was created with manual annotations of the location of the elements in the text. Using the Prodigy framework, the last paragraph of 150 essays that received a maximum score in C5 were selected and manually labeled, serving as the basis for creating a SpanCategorizer model, which can automatically predict labels for similar features to those previously marked. Then, the earlier model was applied to the extended Essay-BR text corpus and stored for analysis. Subsequently, these labels were transformed from spans (SpanCategorizer) into entities (EntityRecognizer) to be quantified using methods from the spaCy library. Furthermore, we adopted a second approach that involved replicating the label recognition model with a stratified database containing more than 100 essays with different scores.

4 Proposal Evaluation

The dataset used in this study was the Essay-BR-Estendido ("Essay-BR" for simplicity) [10], which has been widely used by related work [11,13]. Essay-BR comprises 6,579 essays written in the format of the ENEM on 151 socially relevant topics for Brazil, including fake news, deforestation, and healthcare [10]. Its evaluation process employed by human assessors follows the ENEM format, providing a score for each competency. Only the evaluations for C5 were considered in this work. Additionally, 14 duplicate or empty (string) essays were identified and removed, resulting in 6,563 remaining essays.

To assess the proposed feature extractors, we compared our approach to TF-IDF and Coh-Metrix, two well-established, widely used techniques [9]. In this work, the scikit-learn implementation of TF-IDF was used, with the parameters ngram_range set to (1,2) (i.e., unigrams and bigrams) and min_df = 0.01 (only terms that appear in at least 1% of the essays)[7]. Vectorization was trained on only 80% of the data to prevent potential overfitting and enable a fair evaluation. Concerning NILC, we employed a combination and extension of Coh-Metrix [6] and NILC-Metrix[8] (although not solely from NILC-Metrix, for simplicity, this group is referred to in this work simply as "NILC").

Similar to related work [13], we explored classification and regression models. The following algorithms were employed to create models for predicting C5 scores, selected due to their established use in prior research: Logistic Regression, Random Forest, ExtraTrees, and SVM from scikit-learn, eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM). Importantly, despite Marinho et al. [11] presenting a predefined data split, this work opted for traditional cross-validation with five stratified subsets (maintaining the

[7] These values were defined empirically.

[8] http://fw.nilc.icmc.usp.br:23380/metrixdoc.

proportion of each class). When the problem is converted to regression, stratification occurs before the conversion to preserve the class distribution.

Similarly, the metrics used to evaluate model performance were selected based on related work [11,13]. We used Root Mean Squared Error (RMSE) and Pearson's Correlation for regression. For classification, we used the F1-score and Cohen's Kappa in two versions: linear and quadratic, as suggested in recent studies on the best way to report Kappa [15]. This is necessary because the linear Kappa has a well-established interpretability [8], while the quadratic Kappa enables complementary interpretation by penalizing disagreements differently between adjacent and distant classes. As Kappa is essentially a classification metric, the outputs of regression models are discretized to the nearest class, following the approach in [11].

In that context, we conducted various experiments to understand how to improve AES for C5. Notably, works focused on automated essay evaluation generally use the entire essay as input for feature extraction or multidimensional modeling. However, according to the official ENEM guidelines, the intervention proposal is often found solely in the conclusion (last paragraph). Therefore, our analysis also investigated how considering the full essay or only its conclusion. Moreover, we also varied individually and in combinations of feature groups. Then, based on the preliminary results, we investigated combining them through feature concatenation, averaging the scoring output, and Stacked Generalization aiming for performance improvements [16].

Furthermore, there are three well-established techniques for model optimization. *Feature selection* was performed on the NILC and TF-IDF groups, which have high dimensionality (325 and +2000, respectively). As some features might be redundant or have no effect, feature selection controls the model's complexity and possibly improve performance. *Class weights* aims to assist with the class imbalance in the dataset's class distribution. To address this, the compute_class_weight function from scikit-learn was employed, which implements a simple heuristic to aid in the classification/regression of minority classes (in the case of this dataset, classes 0, 40, and 200). Finally, the *optimization of 10 hyperparameters* of the selected algorithm was performed using the Optuna library, identifying the best ideal values considering the available dataset. Those techniques are computationally expensive, so we only applied them to the best model (found in previous analyses). Finally, we combined all models, as detailed above.

5 Results

This section presents a series of analyses, as described in Sect. 4, towards improving C5 prediction. Notably, all reported results concern the LGBM algorithm, which consistently outperformed the other algorithms in our experiments.

As a preliminary analysis, we investigated the reliability of our NER feature extraction approach, as shown in Table 1. It demonstrates that the models achieved the best results (F1 between 65% and 71%) in the dataset of conclusions with a maximum score in C5. Furthermore, the entity represented by the *effect*

of the intervention proposal obtained the worst results in both models, suggesting potential ambiguity and/or variability of this entity in the text, making its identification more challenging. Overall, given the promising results of the NER approach, we proceeded to investigate along with the other approaches.

Table 1. Results (%) of the entity identification models

	Maximum Score			Stratified Score		
Entity	Precision	Recall	F1-score	Precision	Recall	F1-score
Action	83.72	62.07	71.29	73.58	48.75	58.65
Effect	77.08	56.92	65.49	31.82	16.09	21.37
Form	90.48	51.35	65.52	50.00	26.79	34.88
Agent	88.46	57.50	69.70	59.62	37.35	45.93

First, we compared the proposed group of features specific to C5 (see Sect. 3) to the baseline approaches: NILC and TF-IDF. Regarding classification (clf), Table 2 demonstrates fair-to-moderate results, given the Linear Kappa (just Kappa for simplicity) and Quadratic Weighted Kappa (QWK) values between 0.3 and 0.57. The well-established TF-IDF stood out, whereas NILC also achieved good results. Notably, the concatenation of C5-specific features achieved a performance close to that of NILC, despite its reduced number of features. Furthermore, Table 2 shows that the classification and regression (reg) results were similar. Hence, although the ultimate goal is classification, the subsequent analysis concerns regression, as it has more granular outputs and yields comparable results.

Table 2. Performance of the classifiers and regressors algorithms for each approach

	Kappa	QWK	F1-score	Pearson	RMSE
Proposed (clf)	0.305	0.423	0.306	–	–
Proposed (reg)	0.291	0.423	–	0.478	45.9
NILC (clf)	0.341	0.470	0.331	–	–
NILC (reg)	0.325	0.476	–	0.527	44.3
TF-IDF (clf)	**0.421**	**0.533**	**0.401**	–	–
TF-IDF (reg)	0.388	0.528	–	**0.574**	**42.6**

Second, informed by ENEM's guidelines, we investigated the impact of considering the entire essay (full) compared to only considering the conclusion paragraph (conclusion). As Table 3 demonstrates, the performance in all three groups is better when considering the entire essay. Despite the guideline's suggestion

that the conclusion would have the highest importance for C5, this finding suggests that the rest of the essay also strongly influences the evaluation of the competency, even if indirectly.

Table 3. Feature extraction: full essay x only conclusion

	Kappa	QWK	Pearson	RMSE
Proposed (full)	0.291	0.423	0.478	45.9
Proposed (conclusion)	0.249	0.358	0.416	47.8
NILC (full)	0.325	0.476	0.527	44.3
NILC (conclusion)	0.264	0.384	0.438	47.2
TF-IDF (full)	**0.388**	**0.528**	**0.574**	**42.6**
TF-IDF (conclusion)	0.304	0.436	0.486	45.8

Third, we investigated the model optimization approaches. Building upon the previous findings (i.e., advantages of LGBM, regression, and complete essay), the feature selection process was applied to the high-dimensional groups. These results are presented in Table 4, which compares the performance before (white background) and after (gray background) the optimization. Regarding predictive performance, we found no difference for the NILC group, while there was a slight improvement for the TF-IDF group. However, for both groups, the benefit of lower computational cost is evident by reducing the number of features from 325 to 100 for NILC and from 2,287 to 584 for TF-IDF.

Table 4. Models' performance after feature selection in NILC and TF-IDF groups of features

	Kappa	QWK	Pearson	RMSE
NILC	0.325	0.476	0.527	44.3
NILC (FS)	0.327	0.475	0.524	44.4
TF-IDF	0.388	0.528	0.574	42.6
TF-IDF (FS)	**0.407**	**0.557**	**0.601**	**41.5**

Next, hyperparameter optimization and class weights were used as a final step in the performance improvement process. Table 5 presents the results before (white background) and after (gray background) applying these techniques. The performance of all metrics in the three groups improved. The metric with the most significant difference was Quadratic Kappa, explained by being chosen to be optimized by Optuna. Hence, we demonstrate the benefits from our model optimization attempts.

Table 5. Models' performance before and after the hyperparameter optimization

	Kappa	QWK	Pearson	RMSE
NILC	0.327	0.475	0.524	44.40
NILC (optimized)	0.351	0.516	0.540	45.00
TF-IDF	0.407	0.557	0.601	41.50
TF-IDF (optimized)	**0.422**	**0.604**	**0.627**	**41.00**
Proposed	0.291	0.423	0.478	45.90
Proposed (optimized)	0.307	0.454	0.488	46.50

Finally, we combined all features based on the previous findings (i.e., LGBM, regression task, and all feature groups). For this step, we combined the output (prediction) of each of the three base models through i) a simple average of the scores, ii) a stacked generalization with Linear Regression at the 2nd level, or iii) Ridge Regression at the 2nd level. Table 6 presents a comparison between the three proposed combinations (before optimization) and the *final model*, which used the best of our previous findings, including optimization, for prediction. Table 6 shows that Stacking yielded significantly better results than simple averaging and comparable to the 2nd-level regressors. Overall, the performance of the final model was significantly better than all the other results, revealing that combining generic and C5-specific features contributes to estimating C5.

Table 6. Results of models' C5 estimation using the combination of all features

	Kappa	QWK	Pearson	RMSE
Average	0.361	0.507	0.585	42.00
Stacking - LR	0.413	0.570	0.611	41.10
Stacking - Ridge	0.410	0.565	0.609	41.20
Final model	**0.456**	**0.614**	**0.649**	**39.50**

6 Discussion

While AES has been widely explored for English, limited research targets Portuguese. Consequently, countries like Brazil have been unattended by the potential of AES. For instance, Brazil annually applies ENEM, a large-scale assessment that features a discursive-argumentative essay that has to be manually scored by up to three evaluators, a task known to be tiresome and error-prone. To mitigate such a disparity of AIED adoption across regions, this paper presents a contextualized approach for AES in the context of ENEM.

Our study revealed four main findings. First, our approach for identifying intervention proposal elements (i.e., action, medium, agent, and effect) achieved F1 scores around 0.7, given the complete essay as the input. Particularly, the results demonstrated action and effect were the elements in which we achieved the best and worst results, respectively. These findings demonstrate the feasibility of identifying such elements through NER, the importance of exploring the complete essay instead of the conclusion only, and elements in which our approach is more and less likely to succeed.

Second, the findings revealed that our approach enhances C5's prediction performance. Whereas NILC and TF-IDF achieved better results than the approach we compared, we found that concatenating the three groups led to the best predictive performance. Thus, this finding demonstrates that exploring contextualized nuances of ENEM contributes to developing AES for scoring C5.

Third, our findings revealed that approaching AES as classification and regression problems yielded similar results. This finding was consistent when investigating the group feature independently and together. We focused our analyses on the regression task, as it provides more detailed outputs compared to classification. Hence, these findings revealed that classification and regression tasks in AES are likely to yield comparable results, while the latter provides more detailed insights.

Lastly, we found that the model optimization techniques contributed to more robust models for AES in the context of ENEM's C5. Feature selection contributed to model simplification, whereas class weighting, hyperparameter tuning, and Stacking helped enhance the model's predictive performance. These findings demonstrate the relevance of adopting such approaches to improve model robustness, both in terms of complexity and performance.

Thus, these findings hold many main implications. First, *addressing disparities in AI adoption.* We highlighted the potential to bridge the gap in AI adoption across regions like Brazil by showcasing the feasibility of a contextualized approach to AES tailored for examinations such as ENEM. Second, *the importance of contextual nuances.* The research highlights the significance of considering contextual nuances in educational assessments, particularly by dissecting ENEM's competencies and extracting C5-specific features. Third, *enhancing evaluation tools.* The demonstrated improvements in prediction performance through model optimization techniques suggest pathways for enhancing the robustness of targeted AES systems. Fourth, *advancing pedagogical approaches.* By advocating for the integration of AI technologies sensitive to diverse linguistic and educational contexts, the findings help foster more equitable systems.

Despite the promising results of this study, its limitations should be considered. The main issue relates to the extended Essay-BR dataset, which is unbalanced and contains inconsistent data. Through manual inspection of random samples, problems were observed in how the web crawler process was conducted, resulting in repeated words and extra annotations that should not have been present. Additionally, the dataset spans a period that includes changes in the ENEM evaluation criteria, potentially complicating the automated evalua-

tion process when considering different sets of criteria. As future work, we recommend: (i) creating a new database to address some of the mentioned inconsistencies, (ii) investigating a more recent and complex approach like BERT in this context, and (iii) exploring a hybrid approach (text-based and feature engineering-based), with (i) being a general approach and (ii) and (iii) focusing again on C5.

References

1. Amorim, E., Veloso, A.: A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In: Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 94–102 (2017)
2. Bai, X., Stede, M.: A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. Int. J. Artif. Intell. Educ. **33**(4), 992–1030 (2023)
3. Bittencourt Júnior, J.A.S., et al.: Avaliação automática de redação em língua portuguesa empregando redes neurais profundas (2020)
4. Camelo, R., Justino, S., Mello, R.: 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, pp. 179–186. SBC, Porto Alegre (2020). https://doi.org/10.5753/cbie.wcbie.2020.179
5. Fonseca, E., Medeiros, I., Kamikawachi, D., Bokan, A.: Automatically grading brazilian student essays. In: Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, 24–26 September 2018, Proceedings 13, pp. 170–179. Springer (2018)
6. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: analysis of text on cohesion and language. Behav. Res. Methods Instrum. Comput. **36**(2), 193–202 (2004). https://doi.org/10.3758/BF03195564
7. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). A redação no Enem 2022: cartilha do participante (2022)
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
9. de Lima, T.B., da Silva, I.L.A., Freitas, E.L.S.X., Mello, R.F.: Avaliação automática de redação: Uma revisão sistemática. Revista Brasileira de Informática na Educação **31**, 205–221 (2023)
10. Marinho, J.C., Anchiêta, R.T., Moura, R.S.: Essay-br: a Brazilian corpus to automatic essay scoring task. J. Inf. Data Manag. **13**(1) (2022)
11. Marinho, J.C., Cordeiro, F., Anchiêta, R.T., Moura, R.S.: Automated essay scoring: an approach based on enem competencies. In: Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional, pp. 49–60. SBC (2022)
12. Nau, J., Haendchen Filho, A., Dazzi, R.L.S.: Identificação e avaliação automática da proposta de intervenção em textos dissertativos-argumentativos: uma revisão sistemática da literatura. In: Anais do Computer on the Beach, pp. 493–501 (2019)
13. Oliveira, H., et al.: Classificaçao ou regressao? avaliando coesao textual em redaçoes no contexto do enem. In: Anais do XXXIV Simpósio Brasileiro de Informática na Educação, pp. 1226–1237. SBC (2023)
14. Santos Júnior, J.J.D., et al.: Modelos e técnicas para melhorar a qualidade da avaliação automática para atividades escritas em língua portuguesa brasileira (2017)

15. Vanbelle, S.: A new interpretation of the weighted kappa coefficients. Psychometrika **81**(2), 399–410 (2016)
16. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)