Exploring NLP and Embedding for Automatic Essay Scoring in the Portuguese

Ruan Carvalho[1(✉)], Lucas Fernandes Lins[1], Luiz Rodrigues[2], Péricles Miranda[1], Hilário Oliveira[4], Thiago Cordeiro[2], Ig Ibert Bittencourt[2,5], Seiji Isotani[5], and Rafael Ferreira Mello[3]

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

Abstract. Automated Essay Scoring (AES) presents a promising solution for enhancing the assessment process in education, particularly in standardized tests like Brazil's National High School Exam (ENEM). However, prior research either focuses on the English language or lacks a nuanced consideration of ENEM's particularities, such as competence-based evaluation. This paper investigates AES for ENEM, focusing on a particular competence (C3), which poses challenges due to its subjective nature and high-class imbalance. Leveraging the Essay-BR corpus, which consists of essays aligned with ENEM standards, this research explores traditional Natural Language Processing (NLP) features, contextual embedding representations extracted from BERT models, and a combination of both. Additionally, class weighting techniques are utilized to address class imbalance issues. Results indicate that models based on LGBM and XGBoost, incorporating BERT Embedding alongside NLP features, and augmented by class weighting, yielded the best performance, besides notable enhancements in minority classes accuracy. By presenting a competency-specific analysis, this study contributes towards optimizing AES for ENEM with a contextualized approach to Brazil.

Keywords: Automatic Essay Scoring · ENEM · BERT

1 Introduction

The National High School Examination (ENEM) is a crucial assessment in Brazil, serving as a benchmark for evaluating students' educational competence after completing the foundational education phase. Furthermore, it plays a pivotal role in enabling many students to pursue higher education opportunities[1].

[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. 228–233, 2024.
https://doi.org/10.1007/978-3-031-64315-6_18

ENEM features a discursive-argumentative essay, wherein students must address a proposed problem. The essay must have up to 30 lines and is assessed based on the five competencies, such as organization (C3) and coherence/cohesion (C5).

Each criteria's scores ranges from 0 (complete lack of mastery) to 200 (excellent mastery). Consequently, ENEM's scores range from 0 to 1000 by summing up the scores for all five competencies. For this, two evaluators review the essays. However, the manual grading process, though indispensable, is recognized for its limitations related to the fatigue evaluators are likely to 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 Automated Essay Scoring (AES). AES might partially automate the essay evaluation to improve the efficiency of evaluators while guaranteeing impartial and coherent grades [7]. Most often, AES relies on Natural Language Processing (NLP) and Machine Learning (ML), where regression and classification models are the primary approaches [6]. Particularly within ENEM's context, the emphasis on feature-based approaches is evident, demonstrating promising results on AES tasks given the full essay [4,5]. However, there is a lack of research addressing the assessment of the five competencies individually.

Note that ENEM's evaluators must attribute a score for each competence. Consequently, AES systems that output an overall score, rather than one for each competence have limited practical contribution. Nevertheless, research in this direction is only emerging. To our best knowledge, a single study addressed AES for ENEM's Competence 3 (C3), and the predictive performance in this metric was the worst among the five competencies [4]. To address that gap, this work focuses on C3 of the ENEM, which assesses the student's ability to select, correlate, organize, and interpret information, facts, opinions, and arguments to defend a particular standpoint. Thus, the present study performs an experimental analysis of different strategies for extracting features from ENEM essays.

This paper contributes to AIED research by exploring how to extend AES benefits to an often underserved region: Brazil. Our research contributes to helping Portuguese-speaking people benefit from AES advantages, focusing on a particular aspect of ENEM, C3, presenting a contextualized approach to address a Brazilian challenge. As a continental-sized country featuring millions of students, manually assessing ENEM essays is a prominent challenge. This paper helps address that gap with a targeted intervention for C3.

2 Method

After a manual corpus analysis, a few duplicate essays were identified and removed from the extended Essay-BR [5]. Consequently, the resulting corpus used in this study comprises 6,565 essays, distributed among grades 0 (n = 185; 2.8%), 40 (n = 164; 2.5%), 80 (n = 1601; 24.4%), 120 (n = 3051; 46.5%), 160 (n = 1374; 20.9%), and 200 (n = 190; 2.9%). This demonstrates the imbalance within the dataset, highlighting the need to address it.

This study considered two feature approaches: traditional NLP-based indicators, commonly employed to assess textual cohesion and coherence, and contextual embedding representations extracted from the BERT model.

NLP-Based Features: For each essay, we computed 236 features, organized into seven distinct groups, similar to prior research [7]. The **descriptive** group represents general aspects of the text: number of words; number of words classified as *stop words*; number of sentences; and average of words per sentence.

Coh-Metrix enables the extraction of numerous discourse-level and linguistic attributes that prove useful in assessing cohesion and coherence. These features are categorized into semantic analysis, connectives, lexical diversity, referential cohesion, and syntactic complexity. Eighty-seven metrics were utilized and adapted to the Portuguese language version [1].

Linguistic Inquiry Word Count (LIWC) is a software that, given an essay, compares each word with a set of dictionaries to compute grammatical, psychological, and social aspects within the text. In this study, we adopted the 2007 version [2] to extract 64 features.

Use of connectives is one way to refer to elements previously mentioned in the text. When effectively employed, it ensures a sound logical connection among the ideas presented in the essay. In this group, 33 connectives were computed, 32 following the criteria outlined in [3], and one overall metric.

Lexical Diversity evaluates the student's vocabulary richness by analyzing the frequency of distinct words in the text. For this, we calculate the ratio of various types of words, such as verbs, nouns, adverbs, and adjectives. Besides, three metrics suggested by Palma et al. [8] are also implemented: Hapax Legomena, Yule's K, and Guiraud's Index. In total, we adopted 15 features in this group.

Readability assesses the difficulty of the text from two aspects: sentence length (with longer sentences indicating more incredible difficulty) and word complexity. This study uses five features of this nature [8].

Sentence Overlapping arises when terms mentioned earlier in the text are reintroduced. Six metrics are computed based on the notion that elements shared among adjacent sentences contribute to referential cohesion. Another two metrics are extracted using Term Frequency-Inverse Document Frequency (TF-IDF) across neighboring sentences.

Thematic Coherence is derived from 20 features related to the similarity between the essay and the motivating text. We employed two strategies to represent the essays and their corresponding motivating texts. One strategy utilizes the traditional Term Frequency-Inverse Document Frequency (TF-IDF), while the other leverages embedding representations extracted from BERTimbau-base [9], the Brazilian Portuguese version of the BERT model. Each representation was employed to calculate cosine similarity. The Levenshtein distance was computed using the "fuzzysearch" library[2] and the Jaccard similarity.

BERT Embedding: This approach encodes the essay and its motivating text using the BERTimbau-base [9] to create a combined representation to capture potential relationships. Each element (essay and motivating text) is transformed

[2] https://github.com/taleinat/fuzzysearch.

into a 768-dimensional vector (the default hidden layer size of the BERT model). These vectors are then concatenated into a single vector with 1,536 dimensions. This combined representation aims to encapsulate both syntactic and semantic aspects of the essays, potentially revealing underlying patterns. Afterward, this enriched vector is fed as input to the machine learning algorithms considered in this work for estimating the C3 grades.

Next, we selected classic and more recent algorithms that have demonstrated high performance across previous AES works [7]: Logistic Regression, Random Forest, Extra Trees, K-Nearest Neighbors (KNN), Adaboost, Catboost, LGBM, and XGBoost, all in both classifier and regressor versions. These models were trained using the features extracted from the text and the Embedding obtained from the essay text and prompt using BERT. We adopted the default configuration settings defined in the libraries for all algorithms. Following related work, these algorithms were assessed according to accuracy, F1-score, Cohen's Kappa (linear), Quadratic Weighted Kappa (QWK), and Pearson Correlation, and Root Mean Square Error (RMSE) to provide a comprehensive evaluation, encompassing both classification and regression, enabling a qualitative analysis of the results and comparisons to related work.

Lastly, we defined the experimental steps. C3 scores are categorical: 0, 40, 80, 120, 160, and 200. While this fits the classifiers with a straightforward encoding (i.e., 0, 1, 2, 3, 4, 5), regressors expect numeric values (e.g., 0, 0.2, 0.4, 0.6, 0.8, 1.0). Note that a regressor might assess new inputs (test sets) with intermediate scores (e.g., 0.43) during the learning process. As such, intermediate values become feasible, enabling the algorithm to express "uncertainty". However, the regressor's final score must fall into one of the six categories. For this, the output of the regressor is multiplied by 5 (the value of the highest class) and rounded. If the original output of the regressor is less than 0 or greater than 5, the final value is mapped to 0 or 5, respectively. We also employed a weighted approach for each essay during model training based on the *compute_class_weight* from the scikit-learn library[3], assigning higher weights to elements of minority classes to help the algorithms handle class imbalance.

3 Results and Discussion

LGBM and XGBoost consistently yielded superior predictive performance than the other algorithms analyzed. Therefore, this section focuses on LGBM and XGBoost for conciseness. First, we compared the algorithms' performance - with no approach for handling class imbalance - in three training scenarios: features only (F), BERT Embedding only (E), and both (F+E). These results are presented in Table 1. The best results in each evaluation measure are highlighted in bold. The table shows that the optimal values across all metrics consistently stem from configurations involving both features and Embedding, despite Embedding only closely trails behind.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html.

Table 1. Results achieved by classifiers (c) and regressors (r) considering the features only (F), BERT Embedding only (E) or both (F+E)

Input	Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
F	LGBM(c)	0.548	0.298	0.341	0.430	0.860	0.470
F	XGB(c)	0.537	0.297	0.327	0.414	0.880	0.448
F	LGBM(r)	0.537	0.295	0.347	0.461	0.841	0.500
F	XGB(r)	0.516	0.289	0.329	0.441	0.882	0.464
E	LGBM(c)	0.582	0.328	0.415	0.502	0.823	0.534
E	XGB(c)	0.589	0.357	0.426	0.511	0.821	0.541
E	LGBM(r)	0.588	0.335	0.449	0.555	0.784	0.585
E	XGB(r)	0.559	0.333	0.420	0.534	0.819	0.554
F+E	LGBM(c)	**0.603**	0.352	0.454	0.540	0.797	0.570
F+E	XGB(c)	**0.603**	**0.388**	0.453	0.537	0.809	0.562
F+E	LGBM(r)	0.596	0.346	**0.473**	**0.586**	**0.767**	**0.611**
F+E	XGB(r)	0.563	0.346	0.433	0.548	0.818	0.563

Second, we investigated the best setup (F+E) with class weighting. These results are shown in Table 2, with values that outperformed those of Table 1 highlighted. The findings demonstrate that, especially for LGBM, class weighting contributed to a slight improvement in predictive performance. For instance, Kappa values increased from ≈0.45 to ≈0.48.

Table 2. Experiments considering (F+E) using class weights

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	**0.605**	**0.411**	**0.481**	0.574	0.799	0.589
XGB(c)	0.604	**0.427**	**0.475**	0.558	0.829	0.567
LGBM(r)	0.546	0.359	0.457	**0.588**	0.823	0.592
XGB(r)	0.502	0.324	0.398	0.531	0.886	0.533

These results highlight the efficacy of pre-trained language models. Both classifiers and regressors trained with data extracted through BERT outperformed those trained solely on NLP-based features and achieved almost comparable outcomes when these features were combined with Embedding. Moreover, we found that class weighting also contributed to predictive performance, with confusion matrices showing important enhancements within the minority classes. Thus, these findings reveal that combining NLP and embedding features, along with class weighting, holds promising potential to improve AES systems for ENEM's C3.

Mainly, these findings contribute to optimizing ENEM's assessment. The integration of BERT Embedding alongside traditional NLP features, augmented by class weighting techniques, not only enhances predictive performance but also holds the potential to alleviate issues of evaluator fatigue and subjectivity inherent in manual grading processes. Additionally, handling class imbalance provides a particular contribution towards deploying such models. A model that makes predictions only for the majority classes cannot be practically applied, and by incorporating weighting mechanisms, we were able to mitigate this bias.

Moreover, by focusing on competency-specific evaluation, the study contributes a nuanced approach to AES that aligns closely with the multifaceted nature of ENEM's assessment framework. This contextualized approach fills a critical gap in existing research and highlights the importance of tailoring AES systems to address the unique needs of the Brazilian educational system. By bridging regional disparities in adopting AI in education and offering targeted interventions for improving pedagogical practices, this paper contributes towards leveraging technology to empower students and educators in Brazil.

References

1. Camelo, R., Justino, S., de Mello, R.F.L.: 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 (2020)
2. Carvalho, F., Rodrigues, R.G., Santos, G., Cruz, P., Ferrari, L., Guedes, G.P.: Evaluating the Brazilian Portuguese version of the 2015 LIWC lexicon with sentiment analysis in social networks. In: Anais do VIII Brazilian Workshop on Social Network Analysis and Mining, pp. 24–34. SBC (2019)
3. Grama, D.F.: Elementos coesivos do português brasileiro em córpus de redações nos moldes do Enem: um estudo para a elaboração da CoTex. Phd thesis, Universidade Federal de Uberlândia (2022)
4. Marinho, J., Cordeiro, F., Anchiêta, R., Moura, R.: Automated essay scoring: an approach based on ENEM competencies. In: Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional, pp. 49–60. SBC, Porto Alegre, RS, Brasil (2022). https://doi.org/10.5753/eniac.2022.227202, https://sol.sbc.org.br/index.php/eniac/article/view/22769
5. 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)
6. 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)
7. Oliveira, H., et al.: Towards explainable prediction of essay cohesion in Portuguese and English. In: LAK23: 13th International Learning Analytics and Knowledge Conference, pp. 509–519 (2023)
8. Palma, D., Atkinson, J.: Coherence-based automatic essay assessment. IEEE Intell. Syst. **33**(5), 26–36 (2018)
9. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20–23 (2020, to appear)