PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays

Rafael Ferreira Mello[a,b] **Hilário Oliveira**[c] **Moésio Wenceslau**[a] **Hyan Batista**[a]
Thiago Cordeiro[d] **Ig Ibert Bittencourt**[d,e] **and Seiji Isotani**[f,e]

[a]Universidade Federal Rural de Pernambuco [b]CESAR
[c]Instituto Federal do Espírito Santo [d]Universidade Federal de Alagoas
[e]Harvard University [f]Universidade de São Paulo
{rafael.mello, moesio.wenceslau}@ufrpe.br hilario.oliveira@ifes.edu.br

Abstract

The PROPOR'24 Competition on Automatic Essay Scoring (AES) of Portuguese Narrative Essays evaluated the performance of four participating systems and three baselines for the task of estimating the individual score of four competencies (Formal Register, Narrative Rhetorical Structure, Thematic Coherence, and Cohesion) in narrative essays written by mid-school students in Brazil. The corpus comprises 1,235 essays, divided into train, validation and test sets, and the competitors were evaluated using Cohen's Kappa coefficient and the weighted F1-score. Most submitted systems and baselines leveraged pre-trained language models, particularly Albertina and BERTimbau, demonstrating fair to moderate agreement with human evaluator scores based on the Kappa coefficient. These results highlight the challenge of AES for Portuguese narrative essays while demonstrating the promise of pre-trained language models for future improvement.

1 Introduction

Integrating Artificial Intelligence (AI) into education presents an opportunity to transform the teaching and learning process, offering innovative solutions that enhance efficiency, personalization, and accessibility (Chen et al., 2020). AI holds the potential to impact various educational facets, ranging from content adaptation based on student profiles to delivering personalized, real-time feedback (Cavalcanti et al., 2021). Among the promising applications of AI in education, automatic scoring of textual production, especially essays, stands out (Ferreira-Mello et al., 2019; Chen et al., 2020). AI algorithms can automatically analyze and assess various aspects of an essay, including grammar, cohesion, coherence, argumentative structure, and originality.

Automatic essay scoring (AES) is the task of automatically assigning a grade score to an essay based on a predefined grading rubric (Ramesh and Sanampudi, 2022). The manual correction of essays written by students is a labor-intensive process that places significant demands on teachers and evaluators in terms of time and effort (Costa et al., 2020; de Lima et al., 2023). Moreover, the assessment procedures may be susceptible to individual examiners' personal biases regarding a given topic, resulting in inconsistencies in their evaluations. Developing computer systems capable of automatically evaluating essays based on established criteria can help deal with time demands and consistency challenges in evaluation (Ferreira-Mello et al., 2019). These systems can assist teachers in the classroom by enhancing formative feedback strategies, enabling them to focus on specific areas of writing that require improvement among their students (Ramesh and Sanampudi, 2022).

In recent years, AES systems have experienced advances, particularly in extensively studied languages like English (de Lima et al., 2023). However, progress in low-resource languages like Portuguese still needs to be improved. Most research on Portuguese AES systems concentrates on dissertative-argumentative essays within the high school context (Oliveira et al., 2023a,b), with few studies exploring other domains, such as narrative textual productions commonly utilized in early basic education (Filho et al., 2023).

This shared task aims to contribute to the progression of Portuguese AES systems. In particular, the emphasis is on assessing narrative essays written in Portuguese by students within the Brazilian basic education system. The evaluation was carried out using a corpus comprising 1,235 narrative essays authored by primary school students. Human examiners assessed each essay based on four correction criteria: textual cohesion, thematic coherence, textual typology, and spelling and grammatical errors. The competition involved the participation of four competitors from Brazil and Portu-

gal. Additionally, three commonly used approaches from the literature served as baselines for comparison purposes. The competitors and baselines assessment relied on two automatic evaluation measures: Cohen's Kappa coefficient and weighted F-1 score.

2 Dataset Description

The dataset used in this competition contains 1,235 essays written by students in Brazil's 5th to 9th year of public schools. The students were instructed to write a narrative essay based on a pre-defined prompt given by the teachers. All essays were manually transcribed and anonymized by teachers selected based on their competence with the students in the selected grades.

Afterward, the essays were analyzed by two human evaluators who assessed different aspects of each essay using a pre-defined correction rubric. Given the complexity and subjectivity involved in evaluating this process, disagreements between annotators are common. To mitigate this problem, a third human evaluator with more experience in the task was included to join the annotation team and solve the divergences with the first two annotators. The rubric provides instructive guidance for educators to consider four required competencies:

- **Formal Register:** Appropriate use of the Portuguese language. Aspects such as misspelling words, inadequate use of nominal/verbal agreement and nominal/verbal regency, and inappropriate usage of punctuation symbols are considered.
- **Thematic Coherence:** Adequate understanding of the text production proposal and its development associated with knowledge from different areas, according to the requested proposal, i.e., the plausibility of the text developed concerning the motivating text.
- **Narrative Rhetorical Structure:** Conformity of the text production proposal regarding a Narrative textual typology, articulating ideas, facts, and information in a sequenced and logical way, presenting the constituent elements of this type of textual structure: narrator, place/space, temporal organization, multiple or single characters performing actions.
- **Textual Cohesion:** Correct use of linguistic mechanisms to interconnect text elements, such as words, sentences, and paragraphs.

For each of the four previous competencies, the human evaluators assigned a level ranging from I to V, with **Level I** demonstrating a complete lack of knowledge in the competency domain and **Level V** a complete mastery of the competency.

Figure 1 shows the essay distribution of the full corpus by level for each evaluated competency. The final dataset was divided into three subsets with the following division: 60% (740) for training, 10% (125) for validation, and 30% (370).

The Cohen's Kappa agreement score between annotators 1 and 2 for the four competencies was 0.2475. The overall agreement between the first and third annotators and the second and third annotators was 0.5405 and 0.5650, respectively. Despite the low level of agreement between the first two annotators, it was observed that most of the disagreements were at adjacent levels. For instance, an annotator assigned level III for the essay, whereas the other set level II or IV. This divergence is considered normal in assessments of textual productions, given the subjectivity of the items in the correction rubrics. The final dataset is available at: https://doi.org/10.34740/kaggle/ds/4464018.

3 Competition Participants

This section describes the approaches adopted by the participants in the competition. Five teams initially registered for the competition, but one team did not submit their system's source code. Consequently, evaluating this system on the private test dataset was not feasible. As a result, the competition proceeded with the following four participants:

AESVoting This approach proposes an ensemble of three classifiers (Random Forest, Gaussian Naive Bayes, and Logistic Regression) with a voting/majority rule. Specifically, four separate models are trained, each dedicated to one competency. As input, these models receive an encoded multidimensional contextual representation of the essay extracted using the BERTimbau (Souza et al., 2020). Additionally, during training, the SMOTE technique (Chawla et al., 2002) is employed to address the effects of imbalanced data.

INESC-ID This approach adopted different pre-trained language models based on the Transformers architecture for the Portuguese language. The authors investigated different versions of Albertina PT-* (Rodrigues et al., 2023) model and the BERTimbau-large architecture (Souza et al.,

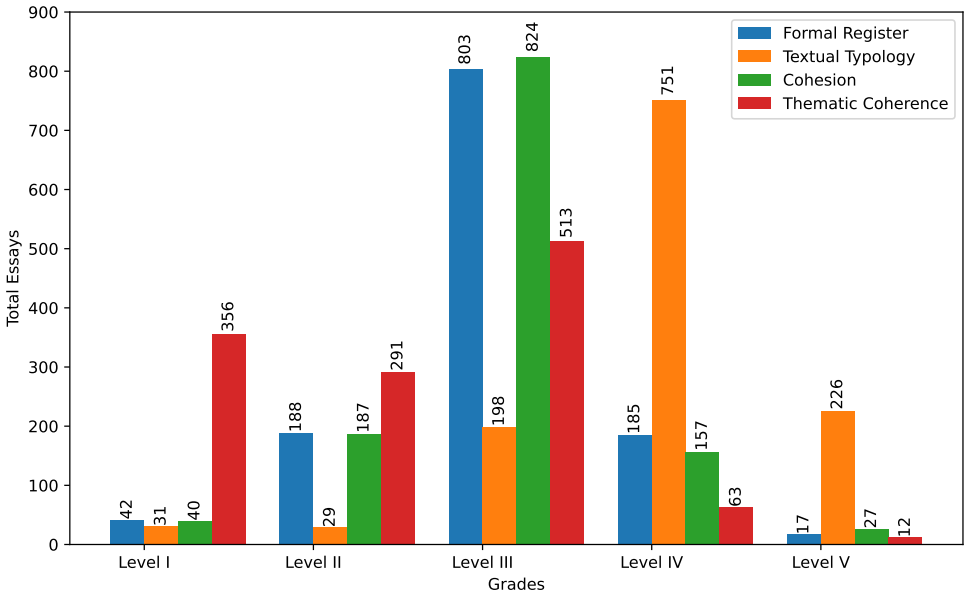

Figure 1: Essays distribution by level for each assessed competency on the complete corpus.

2020). The competitors performed several fine-tuning training steps to estimate the scoring grades of the competencies as a classification problem. The best model for each competency was selected based on the validation set.

PiLN This participant system explored the BERTimbau (base and large) (Souza et al., 2020) with an additional linear layer to predict the scores of the competencies as a regression problem. As the scores predicted by the regression model are continuous values, as post-processing, the estimated scores were rounded to the nearest exact grades (1 to 5). As input, the proposed model receives the essay and the motivating text. The authors tested several configurations and achieved the results using the BERTimbau-large architecture.

Ocean Team In this approach, a two-stage encoding strategy was used as input to various classical machine learning classifiers. First, word embeddings were generated using the BERTimbau model (Souza et al., 2020). Afterward, the Term Frequency – Inverse Document Frequency (TF-IDF) measure was applied to capture term importance. Both representations were used as input to train different classifiers, including Random Forest, XGBoost, Support Vector Machine, AdaBoost, and an ensemble of Extremely Randomized Trees. Four separate models were trained for each competency.

4 Baseline Methods

In addition to the competition participants, the following approaches were evaluated as baselines. These approaches modeled the AES task as a supervised Machine Learning (ML) classification problem and were selected because they present encouraging results in the literature for the AES task in Portuguese in high school essays (Oliveira et al., 2023a,b).

TF-IDF + ML This approach relies on the conventional text representation using the TF-IDF measure. The initial step involved pre-processing the essays by segmenting the text into words, eliminating punctuation symbols, and discarding words with a single character. The words retained after pre-processing from the training corpus were used to build a vocabulary. Each essay was then represented by a vector containing the TF-IDF value for each word in the vocabulary. Finally, these essays were used to train traditional ML algorithms to estimate scores for each competency. The performance of several algorithms available in the scikit-learn[1], eXtreme Gradient Boosting (XGBoost)[2] and Light Gradient Boosting Machine (LGBM)[3], were examined. The algorithm with the best performance for each competence was used as a baseline for

[1] https://scikit-learn.org/
[2] https://github.com/dmlc/xgboost/
[3] https://github.com/Microsoft/LightGBM/

comparison with competing systems.

BERT Embedding + ML This approach involves encoding essays using a multidimensional contextual representation (contextual word embeddings). The representations were obtained through the pre-trained language model of the BERTimbau-base architecture (Souza et al., 2020). The essays were truncated to a maximum of five hundred and twelve (512) tokens, the maximum token sequence length of the BERT model. Each essay is then represented by a vector of seven hundred and sixty-eight (768) values[4], enabling the exploration of syntactic and semantic patterns within the essays. The encoded essays were employed to train traditional ML algorithms for predicting scores across assessed competencies. Then, similar to the TF-IDF approach, several ML models were analyzed, and the top-performing model for each competence was selected as the baseline.

BERT Classifier In this approach, the BERTimbau-base architecture (Souza et al., 2020) was employed, along with an additional dense linear layer, to estimate the scores of the essays' competencies. The model underwent fine-tuning training using the AdamW optimizer with decoupled weight decay and an initial learning rate of $5 * 10^{-5}$. The fine-tuning process was conducted over five training epochs. A BERTimbau model was fine-tuned for each competency considered in the competition.

5 Evaluation Measures

Two automatic evaluation measures commonly used to evaluate AES systems were adopted to assess the baselines and participating competitors (Ramesh and Sanampudi, 2022; de Lima et al., 2023).

Cohen's Kappa coefficient It is a statistical measure used to evaluate the agreement or reliability between two or more annotators or algorithms regarding classifying items into mutually exclusive categories (Cohen, 1960). The Kappa coefficient ranges from -1 to 1, where 1 indicates perfect agreement between, 0 indicates agreement that could be achieved by randomly guessing, and negative values suggest disagreement beyond guessing. The other Kappa values can be interpreted as follows (Landis and Koch, 1977): **(i)** values higher than 0.81 are considered indicative of a very high level of agreement, **(ii)** values between 0.61 and 0.80 suggest a good level of agreement, **(iii)** values between 0.41 and 0.6 indicate moderate agreement, **(iv)** values between 0.21 and 0.4 indicate fair (reasonable) agreement and **(v)** values below 0.21 indicate poor agreement.

Weighted F1-score This evaluation metric is commonly used in machine learning classification problems, especially when significant class imbalances exist. It combines precision and recall metrics into a single score reflecting a model's precision and ability to correctly identify positive examples of each class. This measure computes the weighted average of the traditional F1-score for each class, with weights assigned based on the frequency of each class in the dataset. Therefore, less represented classes have less influence on the overall score, while more represented classes carry higher weight. This metric is particularly valuable in unbalanced multi-class classification scenarios, where simple averaging of the F1-score would not adequately represent the model's effectiveness across all classes. The closer the weighted F1-score value is to 1, the better the model's performance in classifying the different categories.

6 Results

A hold-out strategy was employed to assess both participant competitors and baselines. The dataset was partitioned into three subsets: **training** (60%, 740 essays), **validation** (10%, 125 essays), and **test** (30%, 370 essays) sets. The training and validation sets were provided to competitors to develop their systems, while the test set remained reserved for the final evaluation.

Table 1 shows the results on the test set for each competency based on Cohen's Kappa coefficient and the weighted F-1 score. The competitors shared the source code, which was then used to train and evaluate the system on the blind test set.

The first point to highlight is that the performance of the three baselines remained competitive with the participating systems across all competencies. Specifically, the **TF-IDF + ML** approach for Narrative Rhetorical Structure and the **BERT Embedding + ML** method for cohesion yielded the best results based on the Kappa coefficient values. The **BERT Classifier** demonstrated superior performance regarding the weighted F-1 score in the two previous competitions. The **BERT Embedding + ML** also achieved the best results on the Formal Register competence.

[4]Default representation size defined.

Approach	Cohesion		Formal Register		Narrative Rhetorical Structure		Thematic Coherence	
	Kappa	Weighted F1	Kappa	Weighted F1	Kappa	Weighted F1	Kappa	Weighted F1
AESVoting	0.192	0.567	0.274	0.593	0.219	0.513	0.355	0.552
INESC-ID	0.356	0.691	0.375	0.668	0.284	0.607	**0.548**	0.666
PiLN	0.366	0.692	**0.414**	0.702	0.250	0.616	**0.548**	**0.679**
Ocean Team	0.225	0.647	0.237	0.640	0.187	0.591	0.485	0.621
TF-IDF + ML	0.281	0.650	0.280	0.652	**0.286**	0.623	0.526	0.667
BERT Embedding + ML	**0.367**	0.701	0.407	**0.708**	0.232	0.606	0.448	0.604
BERT Classifier	0.355	**0.702**	0.413	0.704	0.283	**0.626**	0.495	0.643

Table 1: Results of evaluations of participating competitors and baselines in the private test set.

The PiLN achieved the top performance for evaluation measures in Thematic Coherence and the best Kappa coefficient for the Formal Register competency. Also, INESC-ID attained an identical Kappa value to PiLN for Thematic Coherence.

The best outcomes exhibit a reasonable to moderate Kappa coefficient compared to the grades assigned by human evaluators for each competency. These findings show that there is still much room for future progress and highlight the complexity of the task. It is particularly noteworthy the superior performance of participant systems and baselines integrating pre-trained language models such as Albertina and BERTimbau. Such results suggest that leveraging these models holds promise for developing more precise Portuguese AES systems.

References

Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Luciana Costa, Elaine Oliveira, and Alberto Castro Júnior. 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, Porto Alegre, RS, Brasil. SBC.

Tiago Barbosa de Lima, Ingrid Luana Almeida da Silva, Elyda Laisa Soares Xavier Freitas, and Rafael Ferreira Mello. 2023. Avaliação automática de redação: Uma revisão sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.

Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Moésio Silva Filho, André Nascimento, Péricles Miranda, Luiz Rodrigues, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Rafael Mello. 2023. Automated formal register scoring of student narrative essays written in portuguese. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 1–11, Porto Alegre, RS, Brasil. SBC.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023a. Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

Hilário Oliveira, Rafael Ferreira Mello, Péricles Miranda, Bruno Alexandre, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotani. 2023b. Classificação ou regressão? avaliando coesão textual em redações no contexto do enem. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1226–1237. SBC.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.