PiLN at PROPOR: A BERT-Based Strategy for Grading Narrative Essays

Rogério F. de Sousa
IFPI, Picos
rogerio.sousa@ifpi.edu.br

Jeziel C. Marinho
IFMA, Barra do Corda
jeziel.marinho@ifma.edu.br

Francisco A. R. Neto
IFPI, Teresina
farn@ifpi.edu.br

Rafael T. Anchiêta
IFPI, Picos
rta@ifpi.edu.br

Raimundo S. Moura
UFPI, Teresina
rsm@ufpi.edu.br

Abstract

This paper describes the participation of the PiLN team in the PROPOR'24 shared task on Automatic Essay Scoring of Portuguese Narrative Essays. The task aimed to develop methods for automatically evaluating essays to assist teachers in the classroom by enhancing formative feedback strategies, offering a more efficient and cost-effective alternative to human assessment. We approached this task by developing a strategy based on a BERT model; specifically, we fine-tuned a pre-trained BERT model of Portuguese - BERTimbau Large - to calculate scores for each assessed competency, incorporating both the prompt text and the essay text as input. Our simple approach achieved a reasonable result, reaching 4th place with an average score of 0.53985.

1 Introduction

Automated Essay Scoring (AES) is the computer technology that evaluates and scores the written prose (Shermis and Barrera, 2002). It aims to provide computational models for automatically grading essays or with minimal involvement of humans. This research area began with Page (Page, 1966) in 1966 with the Project Essay Grader system, which, according to Ke and Ng (Ke and Ng, 2019) remains since then.

AES is one of the most important educational applications of Natural Language Processing (NLP) (Ke and Ng, 2019; Beigman Klebanov et al., 2016). It encompasses some other fields, such as Cognitive Psychology, Education Measurement, Linguistics, and Written Research (Shermis and Burstein, 2013). Together, they aim to study methods to assist teachers in automatic assessments, providing a cheaper, faster, and more deterministic approach than humans when scoring an essay.

For Portuguese, this area has gained the attention of the community for grading ENEM essays due to publicly available corpora (Marinho et al., 2021, 2022a). ENEM (High School National Exam - *Exame Nacional do Ensino Médio*) consists of an objective assessment and an essay test. The latter comprises a topic (prompt), usually a current problem in Brazilian society, and requires an intervention proposal from the participants. Besides, the text must be written in the argumentative type and not exceeding thirty lines. To grade an essay, ENEM adopts five specific traits that analyze different aspects of an essay[1].

Unlike the ENEM essays, this shared task adopted narrative-type essays and four traits (competencies): formal register, thematic coherence, narrative rhetorical structure, and cohesion. The objective was to develop a computational system capable of estimating a grade for an input essay for each specified trait of interest following the established grading rubric. The task used the average between the weighted F1 score and Cohen's Kappa score, which are widely used in the literature for this task. To deal with this task, we developed a strategy based on BERT (Devlin et al., 2019); specifically, we fine-tuned the BERTimbau model (Souza et al., 2020) for predicting the four traits of narrative essays. With this strategy, we achieved 0.53985 on average and ranked 4th.

The rest of the paper is organized as follows. Section 2 describes the corpus of the shared task. In Section 3, we detail the developed approach and learned lessons. In Section 4, we present the achieved results. Finally, Section 5 outlines the limitations and future work.

2 Corpus

The dataset in this competition contains 1,235 essays written by students in Brazil's 5th to 9th year of public schools. The students were instructed

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

to write a narrative essay based on a motivating text. All essays were manually digitized and anonymized. Afterward, the essays were analyzed by two human evaluators, who assessed different aspects of the essay based on a pre-defined correction rubric. This rubric provides instructive guidance for educators to consider four required competencies: Formal Register, Thematic Coherence, Narrative Rhetorical Structure, and Cohesion. Each dimension was assessed using integer levels ranging from 1 to 5, with higher levels indicating better text quality and language proficiency and lower levels demonstrating a lack of proficiency.

For that, this task made a training set and testing set available. The first one has 740 samples, while the second has 135 samples, where each row in the files contains the essay and a score for each competency. The final ranking was decided based on a blind testing set with 370 essays, according to Figure 1.

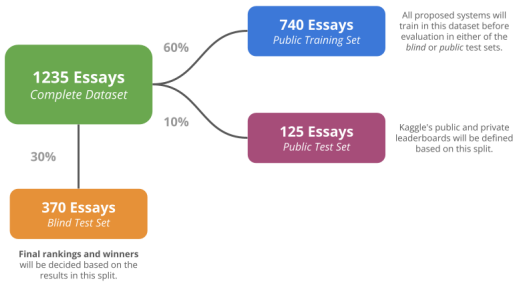


Figure 1: Corpus of the shared task.

In what follows, we detail our proposed strategy.

3 Proposed Method

The adopted method to address this task is based on the work of Matsuoka (2023), which uses a BERT model to evaluate ENEM essays, with some adaptations. The author developed a specialized model called *BERT_ENEM_Regression*, specifically designed for regression tasks, including evaluating essays according to the five competencies established by ENEM. Their study suggested employing the pre-trained Portuguese language model BERTimbau Base (Souza et al., 2020), and they enhanced Marinho et al. (2022a) findings by incorporating essay prompt text as input with the essay in the modeling process.

We employed the same strategy of incorporating the prompt text alongside the essay text as input to the method. However, after testing several configurations, including adjusting the pre-trained model size (Base or Large) and varying the parameters for the fine-tuning process, the results improved when employing the pre-trained BERTimbau Large model. The optimal parameter set used for fine-tuning the model is presented in Table 1.

Parameter	Value
Dropout Layer	0.4
Linear Layer	(1024,4)
Epochs	6
Batch Size	8
Learning Rate	4×10^{-5}
Optimizer	AdamW
Loss	MSE

Table 1: Training parameters for the BERTimbau Large model.

The training corpus is divided into two parts to support the training process and, consequently, the discovery of the best parameters, with 90% for training and 10% for validation/development. The validation set selects the best model during the model fine-tuning process. We chose the model with the lowest loss.

It is worth noting that, as shown in Table 1, we employed a linear layer to make predictions. This layer takes as input, which aligns with BERT's hidden size (1024 inputs) and has an output size of 4. This directly correlates to the four competencies in the essays, with each value corresponding to the score of the respective competency, enabling precise score predictions for each essay.

We also developed a hybrid model, using the features set of the (Marinho et al., 2022b). Although this model improved the results, we decided not to submit it to the shared task.

4 Results

We evaluate our method on the public test set of the shared task. The results for each competency are detailed in Table 2, where FR is Formal Register, TC is Thematic Coherence, NRS is Narrative Rhetorical Structure, and Co is Cohesion. As we can see, the best results are from the F-score metric; the best average was in the second competency, Thematic Coherence.

For the blind test set, we achieved 0.539 in the average between F-score and Kappa metrics. Ta-

Competency	F-score	Kappa	Avg
FR	0.68	0.45	0.56
TC	0.65	0.53	0.59
NRS	0.54	0.19	0.36
Co	0.66	0.34	0.50
Average	0.50		

Table 2: Results for each competency in the public test set.

ble 3 presents the final results available from the shared task organizers. One can see that our approach was ranked 4th.

Team	Score
INESC-ID	0.61
nlpr	0.55
Baseline - BERT Classifier	0.54
Ours (PiLN)	0.539
Baseline - BERT Embeddings+DT	0.532
Tiago de Lima	0.51
Ocean Team	0.46
Baseline - TFIDF	0.44

Table 3: Final result of the shared task.

To better understand the behavior of the model, we generated the confusion matrix for all four competencies, and upon observing it, we can highlight the following insights. Figure 2 presents all matrices. Concerning the Formal Register competency (Figure 2a), the model appears to encounter difficulties distinguishing intermediate scores, as a significant number of essays with intermediate scores mistakenly classified as score 1. Additionally, no essay received the maximum score, indicating that the model also faces challenges recognizing features corresponding to high-quality formal writing. For Thematic Coherence (Figure 2b), it is noticeable that there is a predominance of correct classifications for the lowest score, but on the other hand, almost no essays were classified with the highest score. This suggests that the model may recognize a lack of thematic coherence but has a limited ability to identify more sophisticated thematic coherence.

In Narrative Rhetorical Structure (Figure 2c), the confusion matrix reveals challenges distinguishing between scores 3 and 4, indicating difficulty recognizing excellent narrative structure. Furthermore, there is a relatively high frequency of essays, with the minimum score being classified as higher scores (3) than they should deserve, indicating the need for improvements. Finally, regarding cohesion (Figure 2d), the model appears to have a good ability to distinguish between essays with intermediate levels of cohesion, but, on the other hand, it can be observed that there is a low quantity of essays classified correctly with the maximum score, indicating limitations in identifying advanced cohesion elements.

The source code is publicly available at: `https://github.com/lplnufpi/aes-propor`.

5 Limitations and Future Works

Although our approach reached good results, there is still room for improvement. For example, an error analysis would help to understand why the result in the Narrative Rhetorical Structure was not good. Moreover, a statistical analysis of the essay texts would show insights for incorporating other features into the BERT model.

For future work, we intend to use large language models (e.g., Albertina (Rodrigues et al., 2023)) as data augmentation to balance the corpus.

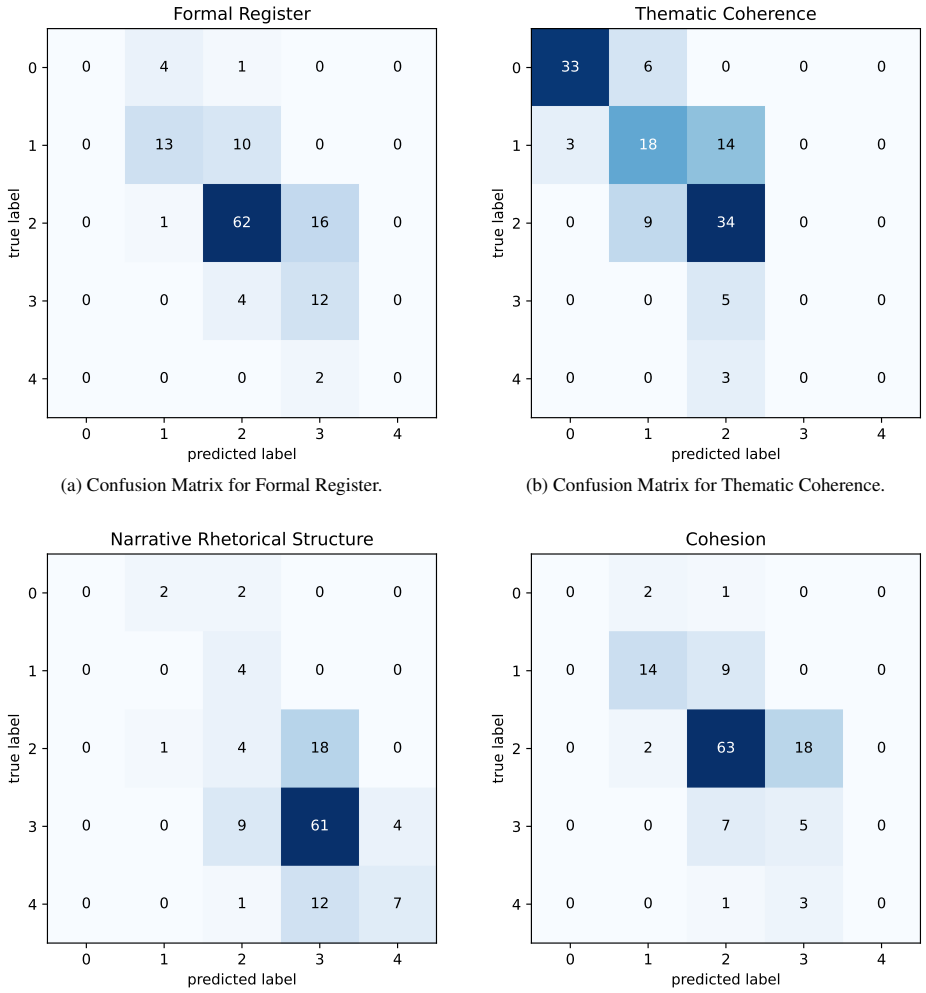
References

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72, San Diego, CA. Association for Computational Linguistics.

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

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308, Macao, China. AAAI Press.

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



(a) Confusion Matrix for Formal Register.

(b) Confusion Matrix for Thematic Coherence.

(c) Confusion Matrix for Narrative Rhetorical Structure.

(d) Confusion Matrix for Cohesion.

Figure 2: Confusion Matrix for Each Competency.

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

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

Felipe Akio Matsuoka. 2023. Automatic essay scoring in a brazilian scenario.

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

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-*.

Mark D Shermis and Felicia D Barrera. 2002. Exit assessments: Evaluating writing ability through automated essay scoring. In *Annual Meeting of the American Educational Research Association*, pages 1–30, New Orleans, LA. ERIC.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, USA.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.