Prediction of Essay Cohesion in Portuguese Based on Item Response Theory in Machine Learning

Bruno Alexandre Barreiros Rosa[1(✉)], Hilário Oliveira[2(✉)], and Rafael Ferreira Mello[1,3(✉)]

[1] C.E.S.A.R School, Recife, PE, Brazil
babr@cesar.school
[2] Federal Institute of Espírito Santo (IFES), Serra, ES, Brazil
hilario.oliveira@ifes.edu.br
[3] Federal Rural University of Pernambuco (UFRPE), Recife, PE, Brazil
rafael.mello@ufrpe.br

Abstract. The essay is considered a useful mechanism for evaluating learning outcomes in writing. Essay correction is a manual task that presents difficulties related to time, cost, reliability, and the subjectivity of the examiner. Cohesion is a fundamental aspect of the text, as it helps to establish a meaningful relationship between its different parts. The automated scoring of cohesion in essays presents a challenge in the field of artificial intelligence in education. This is primarily due to the fact that machine learning algorithms, commonly employed for text evaluation, often overlook the unique characteristics of individual instances within the analyzed corpus. To address this issue, item response theory can be adapted to the machine learning context. This adaptation involves characterizing aspects such as ability, difficulty, discrimination, and guessing in the utilized models. This research aims to analyze the performance of cohesion score prediction in Brazilian basic education essays, using item response theory to estimate the scores generated by machine learning models. The research extracted 325 linguistic features and treated it as a regression problem. Initial results indicate that the proposed approach has the potential to outperform conventional models. The research presents a promising avenue for a more precise evaluation of cohesion in educational essays.

Keywords: Automated essay scoring · textual cohesion · natural language processing · item response theory

1 Introduction

Writing texts is a fundamental skill for an individual to succeed in academic, corporate, or even social life [6]. The production of descriptive, dissertative, injunctive, or narrative essays requires the correct use of linguistic mechanisms

A. M. Olney et al. (Eds.): AIED 2024 Workshops, CCIS 2151, pp. 388–394, 2024.
https://doi.org/10.1007/978-3-031-64312-5_48

essential for the development of writing [16]. This rigor becomes even more crucial when we consider that the evaluation of an essay is a subjective process that involves several criteria [6]. Textual cohesion is one of the most important criteria in this assessment [7–9,13,14]. Cohesion is an indispensable aspect for providing grammatical links and connections between text elements such as words, sentences, and phrases [7]. A cohesive text presents interconnected ideas, which allows the reader to follow the writer's reasoning fluidly [8]. Cohesion is achieved through the appropriate use of linguistic mechanisms necessary for textual construction [7]. In a text devoid of cohesion, the fundamental ideas may be present, but the lack of connections compromises the clarity and effectiveness of communication, making it difficult for the reader to understand and interpret [8]. Therefore, it is expected that improving cohesion in essays will benefit the overall quality of other aspects of writing [1,8].

The use of Automated Essay Scoring (AES) methods brings several advantages, such as reducing the time and costs involved in the correction process, together with minimizing possible biases and human errors [15]. However, when it comes to assessing cohesion in essays, the methods still have limitations. For example, they often fail to capture the reference and sequence of semantic relationships, as well as the logical interconnection between the different parts of the text [1]. In addition, identifying elements of cohesion often requires a contextual understanding that the methods cannot fully replicate [13]. There is also the challenge of correctly interpreting cohesive elements that can have multiple meanings based on the context in which they are used [8]. This makes the task of automatically scoring cohesion in essays an open problem [1,9,13,14].

In English, promising approaches to cohesion have been proposed [1]. In Portuguese, it is still a challenge to treat cohesive elements in essays in an automatic way [9,13,14]. One of the possibilities for dealing with these limitations is to create a committee of estimators that can be selected to perform the scoring in specific contexts. This approach is reinforced by the recurring practice of using techniques to evaluate the performance of Machine Learning (ML) algorithms in order to select the most suitable ones for the various challenges inherent in their application. These evaluation techniques seek to understand the advantages and limitations of the algorithms. Recent studies have taken a different approach to evaluation, in which algorithm performance is evaluated according to the instance level using Item Response Theory (IRT) [12,17].

In this respect, IRT and ML are integrated in a complementary way. The IRT serves as an essential statistical tool. It is used to measure the probability of a respondent answering a specific item correctly based on their latent ability [2]. In this integration, the instances of a dataset in ML can be equated to the items of a test in IRT, while the ML algorithms act as the respondents whose abilities are being assessed. Recent studies have linked this integration and produced promising results [12,17]. However, given a challenging and recently explored context in the scientific field, namely the evaluation of cohesion in essays using ML and IRT models combined in regression problems, this research aims to answer the research question below: *How to predict the cohesion score in essays*

to support teacher correction using item response theory to estimate the scores generated by machine learning models?

To answer this research question, the general and specific objectives were established, as described below:

General Objective

- Analyze the performance in predicting the cohesion score in essays to support teacher correction, using item response theory to estimate the scores generated by machine learning models.

Specific Objectives

- Relate the works that discuss automated essay scoring and item response theory;
- Extract characteristics from essays to run machine learning models for scoring cohesion levels in a real corpus of Portuguese-language essays;
- Propose an algorithm that uses different machine learning models in regression problems to evaluate the level of cohesion in Brazilian basic education essays using item response theory;
- Evaluate the proposed model in relation to traditional machine learning approaches.

2 Background

Recent research has examined the relationship between computationally extracted writing features and human evaluations of cohesion in Portuguese writing. For example, [14] conducted an investigation using 151 features that encompass aspects such as the use of connectors, lexical diversity, readability, and similarity between adjacent sentences, along with several features extracted from the Coh-Metrix tool. The authors compared various regression algorithms to estimate the cohesion score using the Essay-BR database [10]. Following the same research line, the study developed in [13] explored regression algorithms through an attribute-based approach and the BERT language model to estimate scores related to textual cohesion in Portuguese and English. Additionally, explainability methods were employed to provide interpretations of the decisions made by the models for the estimated scores.

The integration between IRT and ML aims to measure the ability of ML models and the difficulty of learning from datasets, showing promising results in recent work [12,17]. An example is the work proposed by [12] that compares the use of IRT across different regressors and provides a theoretical analysis of its parameters and abilities in ML models. This paper models absolute error as a function based on IRT, following a Gamma distribution (Γ), to deal with unlimited positive responses. The research was limited to analyzing the parameters of the proposed Γ-IRT model applied to answers to open-ended Statistics exam

questions, without exploring other writing assessment contexts or extensions of IRT approaches.

In the research [17], the authors proposed a method that applies IRT to evaluate the characteristics of the scores assigned by the AES models and integrates these scores to generate an estimated final score. The study demonstrated that the proposed method, using the IRT Generalized Many-Facet Rasch Model (GMFRM), achieved a higher average precision (QWK 0.7562) compared to individual AES models (QWK 0.7209) and conventional integration methods (QWK 0.7395). This result demonstrates that integrating multiple AES models using IRT can significantly improve performance, evidencing its potential. However, the authors limited themselves to testing a restricted collection of AES and IRT models without cross-validation analysis and only applied the method to one English corpus.

Thus, the use of IRT to analyze and adjust ML models enables more research that can improve the precision and reliability of automated cohesion evaluations in essays. The study by [12] compares the use of IRT in different regressors and provides a theoretical analysis of their parameters and abilities in ML models. [17] focused on integrating the results of AES models using IRT. However, in contrast to previous proposals, this research uses IRT to adjust the output prediction of ML models to produce a new approach for predicting cohesion scores.

3 Method

This study intends to treat the analysis of textual cohesion of essays written in Portuguese as a regression ML problem. Additionally, it plans to integrate IRT to improve score prediction according to the ability of each algorithm. To achieve the proposed objectives, this experimental research will be developed using mixed methods through the following stages:

1. **Data Description:** This stage aims to describe the database to be processed. This research intends to use a Portuguese dataset that includes university-level essays from the Essay-BR extended corpus collected by [10]. This corpus has essays written following the same style as the textual production test of the National High School Exam (ENEM) and includes scores for five competencies used to assess writing, including cohesion [13,14]. There are 6,579 essays, divided into 151 topics, collected from December 2015 to August 2021. These essays were written by high school students, respecting the stipulated limit of a minimum of 8 and a maximum of 30 lines. At least two experts evaluated the essays, and the final score for each competency is the arithmetic mean. In this research, the focus will be on Competency IV, which refers exclusively to textual cohesion in essays;
2. **Feature Extraction:** In this step, the aim is to convert texts into feature vectors for ML models processing while preserving the text's original meaning. These features have previously been used to analyze the rhetorical structure of essays [5], analyze online discussions [11] and evaluate cohesion [13,14].

The intention is to extract measures of linguistic features that include tools such as NILC-Metrix, Coh-Metrix, LIWC, and connective elements. These approaches help to identify, for example, semantic cohesion, referential cohesion, syntactic complexity, lexical diversity, textual simplicity, and readability. In total, 325 measures of linguistic characteristics will be considered to develop the cohesion automated scoring model;

3. **Machine Learning Model Processing:** In this stage, the purpose is to select, train, validate, and test different regression algorithms to estimate essay cohesion scores. For this prediction task, we selected algorithms based on the literature [4,13] that use different approaches, including statistical, decision trees, classical neural networks, Bayesian models, and ensembles. The ML algorithms we intend to adopt in the experiment are: *Bayesian Ridge*, *CatBoost Regressor*, *Decision Tree Regressor*, *Extra Trees Regressor*, *LGBM Regressor*, *Linear Regression*, *MLP Regressor*, *Random Forest*, *SVR*, and *XGB Regressor*. We also plan to use a conventional ensemble of ML model integrations, such as *Stacking* and *Voting* regressors;
4. **Item Response Theory Processing:** In this step, the aim is to predict the score for the level of cohesion assessed in the essays. The IRT calculates the parameters of ability, difficulty, and discrimination of the ML models from the training validation data. Thus, we plan to employ the BIRT-GD library proposed in [3] to calculate the error expectation per model and cohesion level. Subsequently, the ML models will be classified based on cohesion level, so the ML models with the lowest error expectation obtained by IRT are chosen. Finally, the aim is to adjust the output cohesion score by identifying the model instance and cohesion level with the best ability;
5. **Evaluating Results:** In this stage, we will use the evaluation process recommended in the literature [13,14] to compare the results of the regression models and the IRT approach. To ensure the consistency and integrity of the analysis results, we will adopt a 10-fold stratified cross-validation in combination with the following measures: linear Kappa coefficient, Quadratic Weight Kappa (QWK), and Accuracy. It is important to note that these metrics are traditionally used in classification problems. However, in this context, they also apply to regression problems because the values predicted by the regressors are categorized into predefined score ranges.

4 Concluding Remarks

This research aims to improve the evaluation of cohesion in essays to support the correction of Brazilian basic education teachers. Research advances include the development of a model using IRT parameters in order to improve the prediction of ML models for cohesion scores. The study has a detailed timetable that guides the development and deadlines until the thesis defense, scheduled for December 2024. Proposed experiments include: Firstly, we plan to apply the methodology to dissertative-argumentative essays written by Brazilian high school students; Secondly, we intend to incorporate regression models based on Bidirectional

Encoder Transformer (BERT) representations; Finally, we plan to apply the methodology to narrative essays written by Brazilian elementary school students. It is important to mention that this study does not intend to analyze the practical application of the proposed method. This would include developing a learning analysis tool and evaluating instructor and student satisfaction based on the results of our method. We intend to develop this tool in a future line of research.

References

1. Crossley, S.A., Kyle, K., Dascalu, M.: The tool for the automatic analysis of cohesion 2.0: integrating semantic similarity and text overlap. Behav. Res. Methods **51**(1), 14–27 (2019)
2. Embretson, S.E., Reise, S.P.: Item Response Theory. Psychology Press (2013)
3. Ferreira-Junior, M., Reinaldo, J.T., Neto, E.A.L., Prudencio, R.B., et al.: β^4-IRT: a new β^3-IRT with enhanced discrimination estimation. arXiv preprint (2023)
4. Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C.: Text mining in education. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **9**(6), e1332 (2019)
5. Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., Gasevic, D.: 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, pp. 404–414 (2022)
6. Graham, S.: Changing how writing is taught. Rev. Res. Educ. **43**(1), 277–303 (2019)
7. Halliday, M.A., Hasan, R.: Cohesion in English. Longman (1976)
8. Koch, I.G.V.: A Coesão Textual, vol. 22. São Paulo Contexto (2010)
9. Lima, F., Haendchen Filho, A., Prado, H., Ferneda, E.: Automatic evaluation of textual cohesion in essays. In: 19th International Conference on Computational Linguistics and Intelligent Text Processing (2018)
10. Marinho, J., Anchiêta, R., Moura, R.: Essay-BR: a Brazilian corpus to automatic essay scoring task. J. Inf. Data Manag. **13**(1) (2022)
11. Mello, R.F., Fiorentino, G., Miranda, P., Oliveira, H., Raković, M., Gašević, D.: Towards automatic content analysis of rhetorical structure in Brazilian college entrance essays. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 162–167. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78270-2_29
12. Moraes, J.V., Reinaldo, J.T., Ferreira-Junior, M., Silva Filho, T., Prudêncio, R.B.: Evaluating regression algorithms at the instance level using item response theory. Knowl.-Based Syst. **240**, 108076 (2022)
13. 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)
14. Oliveira, H., et al.: 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, pp. 883–894. SBC (2022)
15. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. **55**(3), 2495–2527 (2022)

16. Travaglia, L.C.: Tipologia textual e ensino de língua. Domínios de Lingu@gem **12**(3), 1336–1400 (2018)
17. Uto, M., Aomi, I., Tsutsumi, E., Ueno, M.: Integration of prediction scores from various automated essay scoring models using item response theory. IEEE Trans. Learn. Technol. **16**(6), 983–1000 (2023)