AESVoting: Automatic Essay Scoring with Bert and Voting Classifiers

Tiago Barbosa de Lima
Departamento de Computação –
Universidade Federal Rural
de Pernambuco
Rua Dom Manuel de Medeiros, s/n,
Dois Irmãos - Recife – PE – Brazil
tiago.blima@ufrpe.br

Elyda Freitas
Departamento de
Sistemas de Informação
Universidade de Pernambuco
Caruaru, PE – Brazil
elyda.freitas@upe.br

Valmir Macario
Departamento de Computação –
Universidade Federal Rural
de Pernambuco
Rua Dom Manuel de Medeiros, s/n,
Dois Irmãos - Recife – PE – Brazil
valmir.macario@upe.br

Abstract

In this work, we explore the use of pre-trained models to extract features to automatic essay-scoring tasks using Multinomial Logistic Regression, Random Forest and Guassian Naïve Bayes. We further utilise instance oversampling to mitigate the scarcity of instances to some classes. The results suggest that the addition of synthetic examples turns the model biased and worsens the final result. Therefore, we make use of a voting classifier to mitigate bias which improves the final overall result.

1 Introduction

In Brazil, the National High School Exam (ENEM) work as an evaluation entry exam for many universities (de Lima et al., 2023). One of the requirements is to write an essay in a dissertative argumentative style as proven by academic proficiency (de Lima et al., 2023). The exam produces a demand for the evaluation of millions of essays which is a manually costly operation every year (de Lima et al., 2023). Besides, Automatic Essay scoring (AES) aims to support this task by automatically attributing a score to a textual production often written by a student (de Lima et al., 2023; Sharma and Goyal, 2020; da Silva Filho et al., 2023). Then, several works propose the automatic correction of those essay styles in the literature using different means like pre-trained machine learning models (de Lima et al., 2023; Akio Matsuoka, 2023). Algorithms used for AES such as Logistic Regression, Naïve Bayes and others rely on feature extraction systems to be used as classifiers in several settings (Rudner and Liang, 2002; Kumar et al., 2019a; Sharma and Goyal, 2020; Ludwig et al., 2021; Kumar et al., 2019b). One method to achieve AES is to extract features automatically using a pre-trained model. The work (Beseiso and Alzahrani, 2020) combines manually generated features with model extract features from BERT and Long Short Term Memory to improve AES.

Bidirectional Encoder Representation (BERT) is a widely used pre-trained encoder-only model in several tasks like sentiment analysis, question answer and others (Souza et al., 2020). The work (Akio Matsuoka, 2023) used the Portuguese version of BERT known as BERTimbau developed by (Souza et al., 2020), to automatically score ENEM essays in different categories such as adherence to them, cohesion and coherence, grammatical correction and others obtaining state of the art results.

Despite all the advantages of AES, current methods developed for the Portuguese language in some cases focus on the dissertative argumentative style. On the other hand, works such as (da Silva Filho et al., 2023) evaluate the narrative essays produced by elementary school students in the aspect of a formal register that measures the correct use of linguistics rules for the students. The results show that is possible to achieve good agreement with one of the annotators showing the potential of the application.

Therefore, we explore the use of BERTimbau to extract features to automatically classify students' essays in narrative written style. We use the features from the BERT model as inputs to Multinomial Logistic Regression, Random Forest and Gaussian Naïve Bayes.

2 Materials and Methods

We used 1,235 essays from elementary school students in Brazil proposed by the PROPOR'24. Each essay is written according to a prompt in a narrative style and any personal information is removed automatically. For the competition the texts are evaluated according to four different aspects: a) formal register which evaluates the grammatical aspects of the texts (da Silva Filho et al., 2023); b) thematic coherence which evaluates the if the written text follows the same theme as the motivation

Table 1: The table shows the distribution of grades according to each rubric.

	Formal Register	Rhetorical Structure	Cohesion	Thematic Coherence
1	27	20	26	204
2	111	13	109	175
3	475	123	484	317
4	116	437	108	39

prompt (da Silva Filho et al., 2023); c) rhetorical structure evaluate the uses of discourse marks by students along the essays (Lu et al., 2023) d) cohesion which evaluate the use of linkers and connective ideas along the essay (Oliveira et al., 2023). In each aspect, the text is evaluated in the range from 1 to 5 where 1 is the worst and 5 is the best possible grade. From all 1,235, essays 740 were used for training, 75 for validation and 370 for final test.

We extracted features using the BERTimbau large model similar to what was proposed by (Beseiso and Alzahrani, 2020), but we didn't increment with any other feature. We classified the features using Logistic Regression, Gaussian Naïve Bayes and Random forest algorithms. Since most of the classes are imbalanced, we decided to use the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to mitigate the problem. Further, we used the model's implementation from sci-kit-learn and Huggingface (Wolf et al., 2020; Pedregosa et al., 2011).

The metrics used were the same used by the competition, which are the weighted average f1-score and Cohen kappa metric from scikit-learn library (Cohen, 1960; Pedregosa et al., 2011). Further, instead of considering only the means for all aspects of evaluation we also highlighted which model performance in each aspect.

3 Experiments

We performed preliminary experiments without considering any pre-processing of the original text. The table 2 shows the results for all rubrics evaluated. The results show the Logistic regression model outperforms the others when considering thematic coherence and rhetorical structure and has the best overall result. Meanwhile, random forest is the best model when considering cohesion and formal register.

In a further analysis of the public scores and private, the voting classifiers achieve better results than their counterparts. The results were 0.495 for the public score 0.486 for a private score for logistic regression, and 0.403 e 0.529 for Random Forest for the same metrics. The voting classifiers achieved better results with 0.517 and 0.509 for private and public scores.

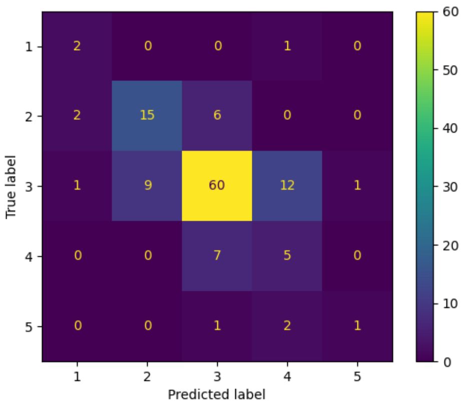

Figure 1: The confusion matrix shows the results of voting classifier for each grade. The expected grade is at the y axis and the predict grade in the x axis.

The confusion matrix showed by the figure 1 suggests that the classifiers performs better on the grades that have more non-synthetic examples.

4 Conclusion and Discussion

In a first analysis, in each rubric the algorithms are bias to produce a lower score as more synthetic data is added. It mostly happens in the rubric of rhetorical structure and cohesion where the imbalance dataset is more evident. Therefore, as more synthetic data is add more bias the models performs what reduces the precision of single model. Furthermore, since the voting class achieves better overall result in the final test set, it turns what that is less bias to the synthetic data of the training set corroborated by the figure 1. Furthermore, the results shows that one of the biggest challenges is to handle imbalance dataset for each rubric in automatic essay scoring task what might be mitigate by the use of Large Language models such as GPT-3 with few shot learning technique (Brown et al., 2020; Touvron et al., 2023).

Table 2: The table shows the result for each model according to each rubric and the result of the voting classifier.

	Formal Register	**Retorical Structure**	**Cohesion**	**Thematic Coherence**
Logistic Regression	0.550	**0.425**	0.480	**0.518**
Random Forest	**0.610**	0.368	**0.482**	0.468
Guassian Naïve Bayes	0.486	0.285	0.384	0.414
Voting Classifier	0.560	0.315	0.472	0.474

References

Felipe Akio Matsuoka. 2023. Automatic essay scoring in a brazilian scenario. *arXiv e-prints*, pages arXiv–2401.

Majdi H. Beseiso and Saleh Alzahrani. 2020. An empirical analysis of bert embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

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

Moésio Wenceslau da Silva Filho, André CA Nascimento, Péricles Miranda, Luiz Rodrigues, Thiago Cordeiro, Seiji Isotani, Ig Ibert Bittencourt, and Rafael Ferreira 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. 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.

A Kumar, P Sharma, and R Singh. 2019a. Ensemble learning approach for predictive modeling using random forest. *Journal of Big Data Analytics in Healthcare*, 4(2):1–11.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019b. Get it scored using autosas—an automated system for scoring short answers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9662–9669.

Hayden Lu, Iel Lykha Dahunog, Jenny Rose Morales, and Norhynie Ranain. 2023. The writing makers of college students: A discourse analysis. *International Journal of Research*, 12(1):15–19.

Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021. Automated essay scoring using transformer models. *Psych*, 3(4):897–915.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Shakshi Sharma and Anjali Goyal. 2020. Automated essay grading: An empirical analysis of ensemble learning techniques. In *Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 2*, pages 343–362. Springer.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.