INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PICOS
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

JOYCE MOURA SILVA

INVESTIGATING METHODS TO DETECT OFF-TOPIC ESSAYS

PICOS, PIAUÍ
2025

JOYCE MOURA SILVA

INVESTIGATING METHODS TO DETECT OFF-TOPIC ESSAYS

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Orientador: Prof. Dr. Rafael Torres Anchiêta

PICOS, PIAUÍ
2025

JOYCE MOURA SILVA

INVESTIGATING METHODS TO DETECT OFF-TOPIC ESSAYS

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Aprovado em 13 de fevereiro de 2025.

BANCA EXAMINADORA:

Prof. Dr. Rafael Torres Anchiêta(Orientador)
Instituto Federal do Piauí (IFPI)

Prof. Dr. Rogério Figueredo de Sousa
Instituto Federal do Piauí (IFPI)

Prof. M^{e}. Manoel Messias Pereira Medeiros
Instituto Federal do Piauí (IFPI)

PICOS, PIAUÍ
2025

INVESTIGATING METHODS TO DETECT OFF-TOPIC ESSAYS

Joyce Moura Silva
Rafael Torres Anchiêta
Rogerio Figueredo de Sousa
Raimundo Santos Moura

ABSTRACT

Automated Essay Scoring is one of the most important educational applications of natural language processing. It helps teachers with automatic assessments, providing a cheaper, faster, and more deterministic approach than humans when scoring essays. Nevertheless, off-topic essays pose challenges in this area, causing an automated grader to overestimate the score of an essay that does not adhere to a proposed topic. Thus, detecting off-topic essays is important for dealing with unrelated text responses to a given topic. This paper explored approaches based on handcrafted features to feed supervised machine-learning algorithms, tuning a BERT model, and prompt engineering with a large language model. We assessed these strategies in a public corpus of Portuguese essays, achieving the best result using a fine-tuned BERT model with a 75% balanced accuracy. Furthermore, this strategy was able to identify low-quality essays.

Keywords: Automated Essay Scoring, Off-Topic Essays, BERT.

1 INTRODUCTION

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

In Brazil, most of the work for the automatic essay grading task focuses on the High School National Exam (ENEM - *Exame Nacional do Ensino Médio*) (CASELI; NUNES, 2024). ENEM is used to assess the quality of high school education and as an admission test for most public and private universities. It provides an essay test and requires the production of a text in the dissertation-argumentative genre on a specific prompt (topic). Each ENEM essay is graded according to five traits (competencies) detailed below.

1. Adherence to the formal written norm of Portuguese.
2. Conform to the argumentative text genre and the proposed topic (prompt) to develop a text using knowledge from different areas.

3. Select, relate, organize, and interpret data and arguments to defend a point of view.

4. Usage of argumentative linguistic structures.

5. Develop an intervention proposal to solve the problem in question.

Although existing works attempt to grade an essay considering the five traits, they do not approach off-topic essays, i.e., essays that do not adhere to the expected prompt (AMORIM; VELOSO, 2017; FONSECA *et al.*, 2018; MARINHO *et al.*, 2022). Off-topic essays are related to the second trait of the ENEM exam, and according to (HIGGINS; BURSTEIN; ATTALI, 2006), they may be classified into two types:

1. **Unexpected topic.** Possibly well-written essays that do not address the expected topic.

2. **Bad-faith.** Essays that mainly consist of text copied from the prompt or with irrelevant musings, such as purposely inserted chunks of text unrelated to the topic and the essay itself.

These essay types are challenging and considered a problem for automated essay scoring systems, as a well-written essay that does not address the proposed topic may receive an overestimated score from an automated grader because of linguistic features, such as text structure and surface (PASSERO *et al.*, 2017). Moreover, (KABRA *et al.*, 2022) has shown that different state-of-the-art automated essay scoring methods fail to provide a low score for poo-quality essays. More than that, off-topic essays occur infrequently, making them difficult to detect automatically. For example, in the 2022 ENEM exam, of the 2,355,395 produced essays, only 31,734 were off-topic, that is, 1.34%[1].

We investigated three strategies for detecting off-topic essays to tackle the above-mentioned challenges. The first is based on handcrafted features for feeding supervised machine learning algorithms to detect off-topic essays, while the second adopts a fine-tuned BERT model. The last approach explores prompt engineering, examining whether the language proficiency retained by a large language model is useful in detecting off-topic essays. These methods were evaluated using the Essay-BR corpus (MARINHO; ANCHIêTA; MOURA, 2021). The fine-tuned BERT model achieved the best result with 75% balanced accuracy.

The rest of the paper is organized as follows: in Section 2, we briefly present related work. Section 3 details the corpus used to evaluate and compare the approaches. In Section 4, we detailed the developed strategies. In Section 5, we reported and

[1] <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>

analyzed the achieved results. Finally, Section **??** concludes the paper by indicating future directions.

2 RELATED WORK

Due to the difficulty of dealing with off-topic essays and the lack of a large corpus of off-topic essays, most works for the Portuguese language focused only on on-topic essays (MARINHO *et al.*, 2022; FONSECA *et al.*, 2018). So, there are a few works on this theme; we briefly discuss them here.

Passero *et al.* (2017) presented a systematic literature review on automatically detecting off-topic essays. They identified five papers, all of them for the English language. The studies dealt with off-topic essays, mainly using textual features, such as Latent Semantic Analysis (LANDAUER; FOLTZ; LAHAM, 1998), n-grams, and Latent Dirichlet Allocation (BLEI; NG; JORDAN, 2003), among others, to classify an essay as on- or off-topic. From this study, the authors found that the approaches had high error rates, and the studies mostly used artificial essay sets for validation.

Passero, Ferreira & Dazzi (2019) adapted the five studies identified by Passero *et al.* (2017) to Portuguese. The authors evaluated the adaptations in a corpus of 2,164 essays[2]. As this corpus has only 12 off-topic essays, the authors used the following strategy to use random on-topic essays as off-topic ones. For each set of N essays of a prompt (negative or on-topic examples), N essays are randomly selected from other prompts (positive or off-topic examples). The adapted work of Klebanov, Flor & Gyawali (2016) achieved the best result. It detects off-topic essays by estimating the topicality of each word in the essay by comparing its occurrence in essays from the same prompt to essays from other prompts.

Pinho, Gaspar & Sassi (2024) compared several supervised machine learning methods to classify an essay as on- or off-topic. To compare the classifiers, they used a private corpus of 1,320 essays, 230 of which were off-topic. A convolutional neural network achieved the best result with 89.4% accuracy with three-fold cross-validation.

In the following section, we present the corpus used in this study.

3 CORPUS

We used the Essay-BR corpus (MARINHO; ANCHIêTA; MOURA, 2021) to evaluate and compare the developed strategies. That corpus contains 4,570 argumentative essays graded according to the five traits of the ENEM exam. Out of the total number of essays, 82 are graded with a score of zero, i.e., 1.79%, a similar value to the number of essays scored with zero in the 2022 ENEM exam. The corpus is organized according to Table 1.

[2] <https://github.com/gpassero/uol-redacoes-xml>

Table 1 – The number of on-topic and off-topic essays in the corpus.

Split	On-topic	Off-topic
Training	3,142	56
Development	671	15
Testing	675	11

To understand why these 82 essays received a zero score, we asked two linguists with experience in grading ENEM essays to evaluate them. For that, we provided the essays and prompts, where each reviewer evaluated all essays independently, following the ENEM criteria[3]. The reviewers agreed that all 82 essays did not follow the expected prompt and received a zero score for not adhering to it.

Table 1 presents the number of on-topic and off-topic essays in the corpus for each set. As we can see, the corpus is very unbalanced, indicating the need to apply sampling strategies to adjust the class distribution of the corpus.

In what follows, we detail our approaches to handling off-topic essay detection.

4 OFF-TOPIC DETECTION STRATEGIES

Due to the unbalancing of the corpus, we first evaluated over- and under-sampling methods. For oversampling strategies, we analyzed approaches for generating synthetic data, such as SMOTE (Synthetic Minority Over-Sampling Technique) (CHAWLA *et al.*, 2002) and ADASYN (Adaptive Synthetic Sampling) (HE *et al.*, 2008). Also, our second oversampling strategy uses random on-topic essays as off-topic ones. We selected 1.543 on-topic essays from the training set and switched their prompts to transform them into off-topic ones. This way, we balanced the training set, resulting in 1.599 on- and off-topic essays. For undersampling, we used a random undersampling strategy from the Imbalanced-learn library (LEMAîTRE; NOGUEIRA; ARIDAS, 2017). It is important to note that these sampling methods were only applied to the training set.

After balancing the corpus, we developed two methods to detect off-topic essays, as depicted in Figure 1. In addition to these methods, we checked whether a large language model may detect off-topic essays. In the following subsections, we detail these approaches.

4.1. FEATURES

We manually extracted six features to detect off-topic essays. We extracted these features by comparing each paragraph of an essay to each paragraph of its correspond-

[3] <https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2023_cartilha_do_participante.pdf>

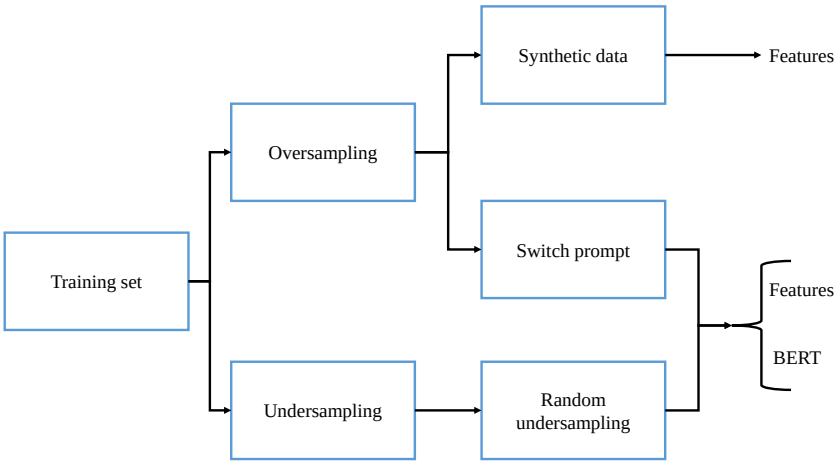

Figure 1 – Approaches to detect off-topic essays.

ing prompt. We then compute the cosine similarity for each compared paragraph pair using the extracted feature values, applying Equation 1, where $\overrightarrow{u} \cdot \overrightarrow{v}$ is the dot product of the two vectors.

$$cos(\theta) = \frac{\overrightarrow{u} \cdot \overrightarrow{v}}{\|\overrightarrow{u}\|\|\overrightarrow{v}\|} \tag{1}$$

Finally, we calculate the arithmetic mean among all cosine similarity values to obtain a similarity between an essay and its prompt. We describe the features in what follows.

Boolean frequency (BF). This feature assigns one value if a word in an essay occurs in its corresponding prompt. Otherwise, a zero value will be assigned, according to Equation 2.

$$BF_{i,j} = 1 \text{ if } i \text{ occurs in } j \text{ and } 0 \text{ otherwise} \tag{2}$$

Term frequency (TF). This feature computes how often a word in an essay occurs in its corresponding prompt. It is the relative frequency of a word in an essay within a prompt, as presented in Equation 3.

$$TF_{i,j} = \frac{f_{i,j}}{\sum_{i' \in j} f_{i',j}} \tag{3}$$

Term Frequency-Inverse Document Frequency (TF-IDF). According to Equation 4, this feature assigns the TF-IDF value to each word in an essay for the analyzed prompt, where $TF_{i,j}$ is the term frequency of i in j, df_i is the number of prompts containing i, and N is the total number of prompts.

$$W_{i,j} = TF_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{4}$$

Cosine of Word Embeddings (COS). To calculate the cosine similarity between an essay paragraph and a prompt paragraph, we got the embeddings for the words, computed the average of the word embeddings for each paragraph, and calculated the cosine similarity between these vectors. We used a 300-dimensional GLoVe model pre-trained for Portuguese to get embedding values (HARTMANN *et al.*, 2017).

Word Mover's Distance (WMD). This feature assesses the distance between two documents even when they have no words in common (KUSNER *et al.*, 2015). It measures the dissimilarity between two text documents as the minimum amount of distance that embedded words of one document need to "travel" to reach the embedded words of another document. It is important to notice that WMD is a distance function, i.e., the lower the distance value is, the more similar the documents are. To get the WMD distance, we first tokenized and removed stopwords of the paragraphs using the Natural Language Toolkit (NLTK) (BIRD; KLEIN; LOPER, 2009); next, we got the embeddings for the words of the paragraphs; finally, to get the WMD distance, we used the method from the Gensim library (ŘEHůřEK; SOJKA, 2010) that receives a paragraph pair encoded as word embeddings as input and returns the WMD value.

Sentence Transformers (ST). It is a modification of the pre-trained BERT network (DEVLIN *et al.*, 2019) that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity (REIMERS; GUREVYCH, 2019). For this feature, we used a multilingual pre-trained model[4] to get the embedding values for each paragraph pair. Then, we computed the cosine similarity between these vectors.

We evaluated some classifiers to assess our approach after the feature extraction step. We used Support Vector Machines (SVM), Naïve Bayes (NB), Decision Tree (DT), MultiLayer Perceptron (MLP), Random Forest (RF) and XGBoost (XGB) from the Scikit-Learn library (PEDREGOSA *et al.*, 2011).

4.2. BERT

We fine-tuned the BERTimbau model (SOUZA; NOGUEIRA; LOTUFO, 2020) to detect off-topic essays. For that, we adopt a method developed by (SOUSA *et al.*, 2024) in which the essays and prompts are used to tune the model. In this way, the model may identify whether an essay adheres to a prompt. We used the large version of the BERTimbau model with the parameters presented in Table 2 to tune the model.

One can see that we employed a linear layer to make predictions. This layer takes the BERT's hidden size (768 inputs) as input and output of size 1 to predict whether the

[4] <https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1>

Table 2 – Training parameters for the BERTimbau large model.

Parameter	Value
Dropout layer	0.4
Linear layer	(1024)
Epochs	6
Batch size	8
Learning rate	4×10^{-5}
Optimizer	AdamW
Loss function	BCEWithLogitsLoss

essay is on- or off-topic. Besides, we used six epochs to train the model and chose the best model through the lowest error in the validation set. We computed the loss using the Binary Cross-Entropy With Logits Loss. This loss function combines a Sigmoid layer and the Binary Cross-Entropy Loss in one class, which is proper for binary classification problems.

4.3. PROMPT ENGINEERING

We investigated whether the language proficiency retained by a large language model may identify an off-topic essay. For that, we used Gemini 1.5 Pro[5], a multimodal large language model developed by Google. This model has demonstrated remarkable capabilities on various tasks and has gained significant attention in diverse domains[6] Furthermore, this model has a free-of-charge option with limits of 15 requests per minute, 1 million tokens per minute, and 1,500 requests per day, which is sufficient for our experiments.

Our prompt design is straightforward. We instructed Gemini to inform whether an essay is on- or off-topic based on a prompt. In this way, we provided delimiters, such as $<$prompt$>$ and “‘ essay ”’ for our inputs to distinguish them, as depicted in Figure 2.

```
Your job is to classify the provided essay as
on-topic or off-topic based on a prompt.

Here is the prompt delimited by <>. <prompt>

Here is the essay that you need to classify,
delimited by triple quotes. “‘essay”’
```

Figure 2 – A simple prompt asking to classify the input essay based on a prompt.

As shown in the above figure, we instructed the model to classify an essay as on- or off-topic given an essay and prompt as input.

[5] <https://gemini.google.com/>

[6] <https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf>

In what follows, we detail the obtained results.

5 RESULTS AND ANALYSIS

We evaluated our approaches using the test set of the Essay-BR corpus (MARINHO; ANCHIêTA; MOURA, 2021). We obtained the best results with the undersampling strategy, as shown in Table 3.

Table 3 – Results with undersampling strategy.

Approach	Class	Precision	Recall	F-score	Balanced Accuracy
Features	on-topic	0.99	0.74	0.85	0.68
	off-topic	0.04	0.64	0.07	
BERTimbau	on-topic	0.99	0.96	0.98	0.75
	off-topic	0.18	0.55	0.27	

Our feature-based approach achieved the best result with the Random Forest classifier, and the fine-tuned BERTimbau model (SOUZA; NOGUEIRA; LOTUFO, 2020) achieved the best overall result. To our surprise, the Gemini model did not identify off-topic essays. As this model reports that it can detect essays that do not adhere to a topic, we expected that it achieved good results for this task.

After assessing the approaches, we performed a feature selection to identify the most important features in our feature-based strategy. To calculate the importance of each feature, we computed the Gini importance (Figure 3), which measures the impurity of a node in a decision tree with a more substantial weight given to the most important features.

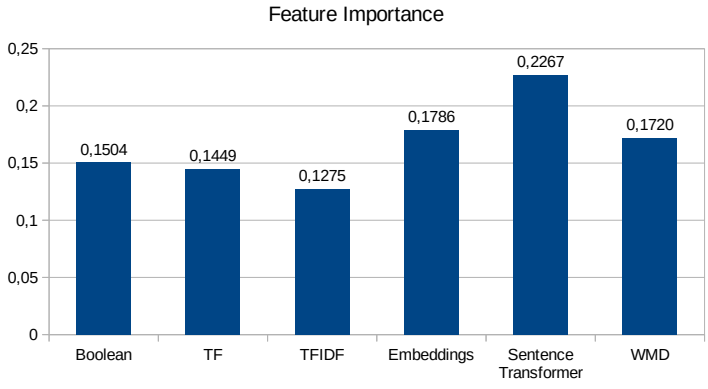

Figure 3 – Importance of each feature.

From this figure, we can see that the TF-IDF contributes the least to the accuracy of the model. At the same time, the features that contribute most to accuracy are the sentence transformer, the cosine of word embeddings, and the word mover's distance.

This result is in line with the study of (HUANG *et al.*, 2023), in which the approach developed by the authors achieved good results in detecting off-topic essays.

Based on the feature selection, we removed the TF-IDF feature and re-evaluated our feature-based approach. With this, our results in recall (off-topic) and balanced accuracy metrics improved a little. On the other hand, the results in the recall (on-topic) and f-score metrics were slightly worse, as shown in Table 4.

Table 4 – Results after feature selection.

Approach	Class	Precision	Recall	F-score	Balanced Accuracy
Features	on-topic	0.99	0.69	0.82	0.70
	off-topic	0.04	0.73	0.07	

We also investigated the results obtained by the BERTimbau model through a confusion matrix, presented in Table 5. From this table, we can see a high number of false negatives (27), justifying the low precision value in the off-topic class.

Table 5 – Confusion matrix of the BERTimbau approach.

		Predicted	
		On-topic	**Off-topic**
Actual	**On-topic**	648	27
	Off-topic	5	6

We further analyzed the false negative predictions, performing an error analysis to understand the misclassifications. With this, we realized that out of twenty-seven false negatives, only four were good essays with scores greater than 800. The remaining essays had grades below 300, with issues such as textual genre (six essays), poor and unstructured text (fifteen essays), and tangent to the topic (two essays).

Our analysis revealed that the strategy of tunning BERTimbau with essays and prompts makes the automatic classification more rigorous than human review. Despite that, this result may be a contribution, as different state-of-the-art automated essay scoring methods fail to provide a low score for low-quality essays (KABRA *et al.*, 2022). However, it is important to analyze the extent to which BERTimbau is strict with poor-quality essays. A deeper inspection of such essays remains necessary for future work.

The developed methods are publicly available at ***https://github.com/liara-ifpi/Off-topic-essay.***

ACKNOWLEDGEMENTS

Primeiramente, agradeço a Deus e à minha família por todo incentivo e apoio, e, em especial, à minha avó, Maria Valdeci, cuja memória jamais será esquecida.

Agradeço ao meu querido orientador, Rafael Anchiêta, pela paciência e pelo tempo dedicados desde o início do curso e durante o desenvolvimento deste trabalho. Sem

sombra de dúvidas, não estaria aqui sem seus ensinamentos. Também agradeço ao professor Rogério Sousa pelos conselhos e pela ajuda durante a realização deste trabalho.

Além disso, agracio às pessoas especiais que essa jornada trouxe para a minha vida, especialmente Arthur Hansel, Daniel Araújo, Gisela Araújo, Maria Cristiana, Marian Karoline e Theodor Almeida.

Muito obrigada a todos.

BIBLIOGRAPHY

AMORIM, E.; VELOSO, A. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In: **Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics**. Valencia, Spain: Association for Computational Linguistics, 2017. p. 94–102. Citado na página 4.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: " O'Reilly Media, Inc.", 2009. Citado na página 8.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 5.

CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. [S.l.]: BPLN, 2024. ISBN 978-65-00-95750-1. Citado na página 3.

CHAWLA, N. V. *et al.* Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002. Citado na página 6.

DEVLIN, J. *et al.* 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Citado na página 8.

FONSECA, E. *et al.* Automatically grading brazilian student essays. In: **Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language**. Canela, Brazil: Springer International Publishing, 2018. p. 170–179. Citado 2 vezes nas páginas 4 and 5.

HARTMANN, N. *et al.* Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. Uberlândia, Brazil: Sociedade Brasileira de Computação, 2017. p. 122–131. Citado na página 8.

HE, H. *et al.* Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE. **2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)**. Hong Kong, 2008. p. 1322–1328. Citado na página 6.

HIGGINS, D.; BURSTEIN, J.; ATTALI, Y. Identifying off-topic student essays without topic-specific training data. **Natural Language Engineering**, Cambridge University Press, v. 12, n. 2, p. 145–159, 2006. Citado na página 4.

HUANG, P. *et al.* A study of sentence-bert based essay off-topic detection. In: **Proceedings of the 4th International Conference on Computing, Networks and Internet of Things**. Xiamen, China: Association for Computing Machinery, 2023. p. 515–519. Citado na página 11.

KABRA, A. *et al.* Evaluation toolkit for robustness testing of automatic essay scoring systems. In: **Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)**. Bangalore, India: Association for Computing Machinery, 2022. p. 90–99. Citado 2 vezes nas páginas 4 and 11.

KE, Z.; NG, V. Automated essay scoring: a survey of the state of the art. In: **Proceedings of the 28th International Joint Conference on Artificial Intelligence**. Macao, China: AAAI Press, 2019. p. 6300–6308. Citado na página 3.

KLEBANOV, B. B.; FLOR, M.; GYAWALI, B. Topicality-based indices for essay scoring. In: **Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications**. San Diego, CA: Association for Computational Linguistics, 2016. p. 63–72. Citado na página 5.

KUSNER, M. *et al.* From word embeddings to document distances. In: **Proceedings of the 32nd International Conference on Machine Learning**. Lille, France: PMLR, 2015. (Proceedings of Machine Learning Research), p. 957–966. Citado na página 8.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. **Discourse processes**, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998. Citado na página 5.

LEMAîTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1–5, 2017. Citado na página 6.

MARINHO, J. C.; ANCHIêTA, R. T.; MOURA, R. S. Essay-br: a brazilian corpus of essays. In: **XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2021**. Online: SBC, 2021. p. 53–64. Citado 3 vezes nas páginas 4, 5, and 10.

MARINHO, J. C. *et al.* Automated essay scoring: An approach based on enem competencies. In: **Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional**. Campinas, Brazil: SBC, 2022. p. 49–60. Citado 2 vezes nas páginas 4 and 5.

PAGE, E. B. The imminence of... grading essays by computer. **The Phi Delta Kappan**, Phi Delta Kappa International, v. 47, n. 5, p. 238–243, 1966. Citado na página 3.

PASSERO, G.; FERREIRA, R.; DAZZI, R. L. S. Off-topic essay detection: A comparative study on the portuguese language. **Revista Brasileira de Informática na Educação**, SBC, v. 27, n. 03, p. 177–190, 2019. Citado na página 5.

PASSERO, G. *et al.* Off-topic essay detection: A systematic review. In: **Proceedings of the XXVIII Brazilian Symposium on Computers in Education**. Recife, Brazil: SBC, 2017. p. 51–60. Citado 2 vezes nas páginas 4 and 5.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 8.

PINHO, C. M. d. A.; GASPAR, M. A.; SASSI, R. J. Aplicação de técnicas de inteligência artificial para classificação de fuga ao tema em redações. **Educação em Revista**, SciELO Brasil, v. 40, p. e39773, 2024. Citado na página 5.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3982–3992. Citado na página 8.

SHERMIS, M. D.; BARRERA, F. D. Exit assessments: Evaluating writing ability through automated essay scoring. In: **Annual Meeting of the American Educational Research Association**. New Orleans, LA: ERIC, 2002. p. 1–30. Citado na página 3.

SOUSA, R. F. de *et al.* PiLN at PROPOR: A BERT-based strategy for grading narrative essays. In: **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2**. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 10–13. Citado na página 8.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: **Proceedings of the 9th Brazilian Conference on Intelligent Systems**. Online: Springer International Publishing, 2020. p. 403–417. Citado 2 vezes nas páginas 8 and 10.

ŘEHůřEK, R.; SOJKA, P. Software framework for topic modelling with large corpora. In: **Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks**. Valletta, Malta: University of Malta, 2010. p. 46–50. Citado na página 8.