The Name of the Title Is Hope

ANONYMOUS AUTHOR(S)*

Machine learning models are susceptible to the dataset used. Dealing with limited or imbalanced datasets is challenging, and a commonly adopted solution is data augmentation. For example, expanding the training set in a computer vision problem may involve rotation and resizing images; however, this task is more complex when dealing with textual data. This work investigates the use of ChatGPT for data augmentation in a dataset of argumentative essay texts from the National High School Exam (ENEM), which are used as selection criteria for entry into public universities in Brazil. Techniques in Natural Language Processing (NLP) are employed to extract 236 metrics from each text and use BERT to assess students' abilities to select, relate, organize, and interpret information, facts, opinions, and arguments supporting a viewpoint. Our results show that the long argumentative essays generated by ChatGPT did not improve the performance of machine learning models. Moreover, ChatGPT could not adequately classify its synthetic data, suggesting poor quality of the generated data, and did not outperform machine learning models in classifying real data.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Information extraction**.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:

Anonymous Author(s). 2018. The Name of the Title Is Hope. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The National High School Examination (ENEM) is one of Brazil's primary assessments to gauge students' educational proficiency after their basic education cycle. It is through this examination that numerous students gain access to higher education in public universities via the Unified Selection System (SISU), as well as in private institutions through the Student Financing Fund (Fies)[1].

The assessment encompasses a series of objective questions spanning four domains of knowledge (languages; codes and their technologies; human sciences and their technologies; natural sciences and their technologies; and mathematics and their technologies) and a discursive-argumentative essay (ENEM essay). In this manner, the exam administrators present a theme and a problematic issue to be resolved or alleviated by the participants within the allocated examination time. In addition, motivational texts are provided that underscore the presence of the problem situation within society [10].

The Enem Essay is a type of dissertative-argumentative text in which the student is required to address a topic and advocate a viewpoint on it, concluding with a proposed intervention to mitigate or resolve the indicated problem. The essay has to be written within a maximum of 30 lines, and the theme sentence typically pertains to a current issue within Brazilian society [10].

The evaluation of the essay is predicated upon five criteria or competencies:

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

Manuscript submitted to ACM

(1) Demonstrating mastery of the formal written mode of the Portuguese language;
(2) Comprehending the essay prompt and applying concepts from various fields of knowledge to elaborate on the topic;
(3) Selecting, correlating, organizing, and interpreting information, facts, opinions, and arguments in defense of a particular standpoint;
(4) Exhibiting an awareness of the linguistic mechanisms necessary for constructing the argumentation;
(5) Devising an intervention proposal for the addressed problem while upholding human rights.

Each of these competencies can assume scores ranging from 0 (complete lack of mastery of the mode) to 200 (excellent mastery of the mode). Consequently, the overall grade for the Enem Essay can vary between 0 and 1000, summing up the score of all the competencies. Upon the conclusion of the assessment period, these texts are appraised by two assessors tasked with evaluating "a standpoint substantiated by consistent and well-structured arguments with coherence and cohesion." In the event of a disparity in scores, a third assessor is required to reevaluate the essay, aiming to establish a potential consensus on the grade[2].

The manual grading process, despite its necessity, is widely acknowledged for its significant drawbacks concerning the fatigue experienced by evaluating assessors due to its repetitive nature. Furthermore, due to its human-centric nature, the grading process is susceptible to numerous inconsistencies and biases, which may result in unreliable assessments. Consequently, a potential resolution to this predicament lies in deploying intelligent computational systems that operate automatically, enhancing assessors' efficiency while ensuring a uniformly unbiased and coherent grading outcome [11, 15].

There are techniques grounded in Natural Language Processing (NLP) that fulfill this role of receiving an essay and automatically delivering the corresponding grade using Machine Learning (ML) models [13–15]. In the literature, this challenge is addressed as Automated Essay Scoring (AES) [9], wherein the significance of regression and classification models is discussed as primary approaches, along with the challenges associated with working with texts in the Portuguese language, as is the case with the ENEM essay.

Although there are many automated essay scoring works in English, there are still few in Brazilian Portuguese (PT-BR) [7]. Another challenge faced in this domain is data imbalance. Generally, PT-BR corpora used in most studies have a major class with 18 times more instances than the minor one [7]. Therefore, estimating the score by competency and, in total, becomes a complicated task. This work focuses on Competence 3 (C3) of the ENEM, which assesses the student's ability to select, correlate, organize, and interpret information, facts, opinions, and arguments to defend a particular standpoint.

Thus, the present study performs an experimental analysis of strategies for estimating C3 considering the extended Essay-BR dataset [12], which comprises 6,579 essays from the ENEM. To address the data imbalance issue, two approaches are combined: 1) assigning weights based on the size of each class and 2) data augmentation. In the first approach, more importance is given to a class with fewer examples during model training. In the second approach, ChatGPT is utilized to generate new essays based on the original essay topics. The goal is to determine if these approaches, individually and in combination, lead to better machine-learning models. Additionally, a thorough evaluation of the adopted LLM is conducted to assess its ability to classify texts compared to the trained models.

The current task of estimating the C3 score has been approached as a classification and regression task. In total, two algorithms were employed in their classification and regression versions. The algorithmic performance was assessed

[2]https://www.gov.br/inep/pt-br/assuntos/noticias/enem/conheca-o-processo-de-correcao-das-redacoes

regarding accuracy, F1-score, Cohen's Kappa, Quadratic Kappa, Pearson Correlation, and Root Mean Square Error (RMSE). The results indicate that using class weights alone improved performance, and it is suggested that the LLM is unsuitable for long texts. The synthetic data generated by the LLM did not enhance model training, and the LLM itself could not estimate the essays' scores from the adopted corpus and the texts generated by the LLM.

2 Related work

In this section, we present related works that employed Large Language Models (LLMs) to address the challenge of imbalanced data.

In [16], a small portion of the training set is utilized to fine-tune the GPT-2 model, enabling it to generate new sentences. An optimization process is conducted using the Monte Carlo Tree Search method to decide which generated sentences would be incorporated into the training set. The authors compared this approach, performed on English texts, with Non-Guided Data Generation (NGDG) and achieved a 5% performance improvement.

In [1], the initial step involves checking whether the text is short (up to 280 characters) or long. In the latter case, the initial words, such as the title, are used as an initial context to guide the creation of long synthetic data. Fine-tuning is performed on GPT-2 by adding tokens before each text to indicate the data class, and the model's temperature parameter is adjusted to introduce uncertainty in generating new sentences. Subsequently, GPT-2 generates new texts, and a filter is applied to increase the probability that the synthetic data maintains the original labels. This methodology was applied to 11 datasets (all in English), with five containing long texts. The method achieved improvements of 2 to 4 points in F1-score in some datasets, while in others, it resulted in a performance similar to the baseline.

In [5], techniques for data augmentation in formative assessments (short texts) for high school education were explored. Four techniques were addressed, including using masks with BERT for word substitution, noise injection, substitution by hyponym/hypernym, and oversampling of existing data. The results demonstrated that data augmentation improved performance, although none of the four techniques studied outperformed the others.

Building upon the previous work, [6] employed generative AI for data augmentation. They used the GPT-3.5 model (text-curie-001) with the command "paraphrase this sentence" alongside an instance of the real dataset. Three values of the model's "temperature" parameter were tested to investigate whether it affected the performance of the final classifier, which is a BERT model. These approaches were compared with a baseline (BERT model without data augmentation), self-augmentation, and a priori model (always choosing the majority class). The data augmentation approaches also outperformed the baseline in all seven datasets studied. However, the a priori model won in two datasets (the most imbalanced ones), the one trained with self-augmentation in one, while the model using GPT-3.5 outperformed in four. It was also observed that higher "temperature" values generated sentences with greater diversity, but this did not alter the maximum performance of models trained with synthetic data from GPT-3.5.

In [8], AugGPT used ChatGPT to rewrite each sentence (short texts) in the training dataset and produce several new sentences while preserving the semantics of the original. The results are compared with 21 other data augmentation methods on three datasets, and the proposed method exhibits better accuracy in all scenarios. The authors also explore an approach where ChatGPT is trained with few-shot learning to classify texts. They concluded that ChatGPT performs well in simple tasks, such as identifying symptoms based on a single sentence description, but struggles in more complex tasks.

This article investigates the automatic correction of argumentative essays from the ENEM in PT-BR considering the extended Essay-BR dataset, which faces the challenge of imbalanced data distribution. For nearly all five assessed

competencies in the ENEM, the extreme scores (the first and second lowest and highest grades) have significantly fewer examples than others.

Inspired by previous research, we leveraged GPT-3.5 to generate essays and determine whether synthetic data can lead to improved models for estimating scores in Competency 3 of the ENEM, specifically to address this issue. Beyond the language aspect, this article distinguishes itself from others due to the length of the texts utilized. On average, each essay comprises 12 sentences and 290 words, making them considerably lengthy (in work by [1], texts are considered short if they contain up to 280 characters). Thus, ChatGPT is applied in a distinct context, and we also assessed its consistency by requesting it to evaluate the generated synthetic data.

2.1 Research Questions

This work aims to answer the following research questions (RQ):

- **RQ1**. Do the synthetic data produced by the ChatGPT improve the machine learning models' performance on estimating the C3?
- **RQ2**. Including the ChatGPT as estimator, what is the best model to estimate the C3 grade?

3 Methodology

3.1 Original Dataset

The dataset used for the experiments described in this article is the extended Essay-BR, a corpus of 6,579 argumentative essays distributed in 151 different topics [12]. The corpus was created to fill the resource gap for developing alternative methods for the automatic assessment of essays in Portuguese. These data consist of multiple essays written by Brazilian high school students on an online platform and evaluated by experts in five competencies. The evaluation process follows the same criteria adopted in the ENEM, the main pathway to access the top universities in Brazil [12].

One of the significant challenges inherent in the Extended Essay-BR is data imbalance. It is common for scores at the extremes to have few examples. Table 1 displays the distribution of essays by C3 scores, which is the focus of this research. As can be seen, scores 0, 40, and 200 each have fewer than two hundred examples, while the majority score has over 3,000. Thus, we adopted LLMs as an augmentation strategy of examples of the minority classes.

Table 1. Distribution of essay grades considering C3.

Grade	# essays
0	185
40	164
80	1,601
120	3,051
160	1,374
200	190

3.2 Feature Extraction

For each essay, features are extracted using natural language processing techniques. Machine learning algorithms used these features to train and predict the C3 score.

In this work, 236 features are computed for each essay from the Portuguese versions of Coh-Metrix [2] and Linguistic Inquiry Word Count (LIWC) [3] tools. Additionally, BERTimbau (neuralmind/bert-base-portuguese-cased) [17] was adopted for extracting 768 embeddings from both the raw essay text and the prompt. Thus, a total of 1,772 different features are extracted from each essay. This methodology was also employed in [4], where it is described in detail.

3.3 Adopted ML Algorithms

Several classical and recent white and black-box algorithms well-known for their high performance in various tasks were selected, including Logistic Regression, Random Forest, Extra Trees, KNN and Adaboost from scikit-learn[3], Catboost [4], LGBM [5], and XGBoost [6], with both classification and regression versions. The algorithms adopted were configured with their default parameterization defined in their libraries. These models were trained using the features extracted from the text and the embeddings obtained from the text and the prompt by BERT (refer to Section 3.2).

3.4 Evaluation Metrics

The set of evaluation metrics, as follows, was chosen to evaluate the algorithms considered in this work properly:

- Accuracy - It represents the percentage of elements in the test set that were classified correctly: $\frac{True\ predictions}{All\ predictions}$.
- F1-score - A classification metric that already encapsulates various information, more comprehensive and informative than accuracy, precision, and recall by themselves: $\frac{2*precision*recall}{precision+recall}$.
- Cohen's Kappa - A metric of agreement between assessments, which measures the agreement/disagreement between two evaluators (in this case, the machine learning algorithm and the original human evaluator). This metric is widely used in related works, as it penalizes randomness more than accuracy, for example: $\frac{p_o-p_e}{1-p_e}$.
- Square Kappa - It is like the previous metric, but with weights, but considering errors between different classes. In the original (linear) Kappa, a disagreement between classes 0 and 1 and 0 and 3 is considered equal. In contrast, in the quadratic version, the disagreement between 0 and 1 is considered much less severe than between 0 and 3.
- Pearson Correlation - A regression metric that indicates how much one continuous variable relates to another continuous variable (how much one "explains" the other): $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$.
- RMSE (Root Mean Square Error) - Unlike the previous metrics (where higher values are better), this metric is the inverse, representing how much the prediction deviates from the expected value: $\sqrt{\Sigma\frac{(\hat{y}_i-y_i)^2}{n}}$.

3.5 Experimental Steps

In this work, we conducted experiments that considered estimating the C3 score as either classification or regression. Originally, the scores were discrete, ranging from 0 to 200, in 6 classes (0, 40, 80, 120, 160, 200), with an associated order. For the classification experiments, we considered the classes [0, 1, 2, 3, 4, 5]. However, by modeling the problem this way, the classes do not necessarily preserve the order among them. Nevertheless, the advantage is that it remains faithful to the original problem, providing one of the six possible outputs for each input essay.

Another possibility is to adapt the problem for regression. In this format, after normalization, the inputs will have only six possible values (0, 0.2, 0.4, 0.6, 0.8, 1.0). However, the algorithm may evaluate new inputs (test sets) with

[3]https://scikit-learn.org/
[4]https://github.com/catboost/catboost/
[5]https://github.com/Microsoft/LightGBM/
[6]https://github.com/dmlc/xgboost/

intermediate scores (e.g., 0.43) during training. This way, the ordering is preserved (e.g., 0.63 is greater than 0.41), and intermediate values are possible, allowing the algorithm to express its "uncertainty" through an intermediate score. Nonetheless, the final score must fall into one of the six categories: [0, 1, 2, 3, 4, 5]. For this purpose, the output of the regressor is multiplied by 5 (the value of the highest class), and then rounding to the nearest integer is applied. If the original output of the regressor is less than 0 or greater than 5, the final value is mapped to 0 or 5, respectively.

4 Results

4.1 RQ1: Do the synthetic data produced by the ChatGPT improve the machine learning models' performance on estimating the C3?

The high imbalance in the extended Essay-BR dataset makes it challenging for algorithms to capture the characteristics of different C3 score levels. Therefore, we will investigate whether increasing synthetic data for the minority classes (scores 0, 40, and 200) helps the learning algorithms achieve better results in estimating C3 (As part of RQ1).

The algorithms described in Section 3.3 had their performances evaluated after data augmentation. However, in the initial tests, LGBM and XGBoost showed the best performance from the perspective of the metrics described in Section 3.4. Thus, the following subsections will only present the results obtained by the best algorithms.

4.1.1 Preliminary data augmentation for the minority classes. Herein, we performed a preliminary analysis of the ChatGPT contribution on data augmentation. In this initial test with the OpenAI library, we used the ChatGPT to generate a sample of synthetic essays of the minority classes. The following dataset was created:

- score 0: 5 essays by theme (755);
- score 40: 4 essays by theme (604);
- score 200: 6 essays by theme (906).

This resulted in generating 2265 new essays. This was done using the *openai.ChatCompletion.create* method with the following instructions:

- a message defining the role of ChatGPT (*role: system*) with the following content: *Você é um aluno prestes a concluir o ensino médio no Brasil.*;
- a message with details for creation of the content (*role: user*): *Crie uma redação nota X na Competência 3 do ENEM com cerca de 200 palavras contendo no máximo 4 parágrafos e com base nos textos motivadores a seguir: TEXTO_MOTIVADOR.*

As it can be seen, even including the 2265 essays of the minority classes, these data instances remain insufficient to achieve a balanced dataset. It was performed to provide preliminary insight into the contribution of ChatGPT as a data augmentation strategy. To alleviate the imbalance in this scenario, we adopted two approaches: one without assigning weights to the classes and another utilizing the *compute_class_weight* function from the *scikit-learn* library. In the second approach, we assigned weights based on the number of examples; if class X has twice as many instances as class Y, the algorithm penalizes an error in class Y two times more. Consequently, the model is inclined to focus more on the minority classes.

The two approaches previously mentioned were applied in two different experiments. The first experiment performs a 5-fold cross-validation methodology but separates synthetic data from real data. In each fold, synthetic data from the minority classes (0, 40, and 200) were incorporated alongside real data for model training. Nonetheless, the essays in the test set consisted of original data. The machine learning algorithms were also compared without adding synthetic data

to the original dataset. The second experiment assesses the machine learning algorithms using a dataset that included all essays (real and those generated by the LLM model) without distinguishing between real and synthetic essays and applied cross-validation.

Experiment separating real and synthetic data. We conducted the first experiment without class weights. The results are presented in Table 2, where we highlight the top two values for each metric. It is important to mention that LGBM(c) and XGB(c) are the algorithms in a classification mode, and the LGBM(r) and XGB(r) in a regression mode. The first column informs if the augmentation was applied or not, and as can be seen, the best results were obtained without using data augmentation. In general, LGBM(c) and LGBM(r) presented the best results in most of the metrics.

Table 2. Results on experiments without class weights for minority classes.

Augmentation	Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
Yes	LGBM(c)	0.600577	0.366783	0.446895	0.526314	0.835121	0.541983
Yes	XGB(c)	0.59738	**0.37498**	0.440407	0.514022	0.847429	0.529212
No	LGBM(c)	**0.603471**	0.351518	**0.453986**	0.539694	**0.797342**	**0.570293**
No	XGB(c)	**0.603167**	**0.387692**	0.453111	0.53724	0.80919	0.56196
Yes	LGBM(r)	0.561595	0.31537	0.420564	0.532653	0.81547	0.555902
Yes	XGB(r)	0.537686	0.333757	0.403732	0.516796	0.860535	0.52686
No	LGBM(r)	0.595554	0.346394	**0.473399**	**0.586191**	**0.766673**	**0.611184**
No	XGB(r)	0.562813	0.346393	0.43297	**0.547695**	0.818452	0.562962

We also applied weights to the essays during training with cross-validation based on the class distribution, as the proportion varies depending on whether synthetic data from scores 0, 40, and 200 are used. The results are presented in Table 3 and those models that did not use augmentation had their previous results improved regarding the ones achieved without class weights (shown in Table 2). Figure 1 depicts the confusion matrices of the LGBM whose models achieved the best results. It can be seen that the use of synthetic data in classes 0, 1, and 5 (scores 0, 40, and 200, respectively) does not necessarily enhance the model's performance in these groups. Additionally, we observe that the models encounter difficulty when dealing with instances at the extremes (classes 0 and 5), as in these cases, misclassifications often result in errors exceeding a distance of 1. When the model misclassifies other classes (1 to 4), it typically assigns an adjacent label. Although we have not presented figures for XGB models, we confirm that they behaved similarly.

Table 3. Results on experiments with class weights and data augmentation for minority classes.

Augmentation	Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
Yes	LGBM(c)	0.59936	0.389017	0.465727	0.539083	0.862744	0.543738
Yes	XGB(c)	0.599817	0.380962	0.457331	0.528333	0.859561	0.535864
No	LGBM(c)	**0.604993**	**0.411095**	**0.481356**	**0.574442**	**0.798583**	**0.589434**
No	XGB(c)	**0.603929**	**0.427098**	**0.474674**	0.558426	0.829357	0.567253
Yes	LGBM(r)	0.527943	0.319584	0.422028	0.551708	0.843918	0.559067
Yes	XGB(r)	0.504642	0.330088	0.395734	0.524265	0.892416	0.526897
No	LGBM(r)	0.546369	0.35932	0.456899	**0.588318**	**0.822813**	**0.591777**
No	XGB(r)	0.502055	0.323582	0.39753	0.5307	0.886166	0.532962

Experiments without differentiation between synthetic and real data. In this cross-validation setup, synthetic data is included not only in the training, as performed in the last experiment, but also in the test set. Table 4 shows the

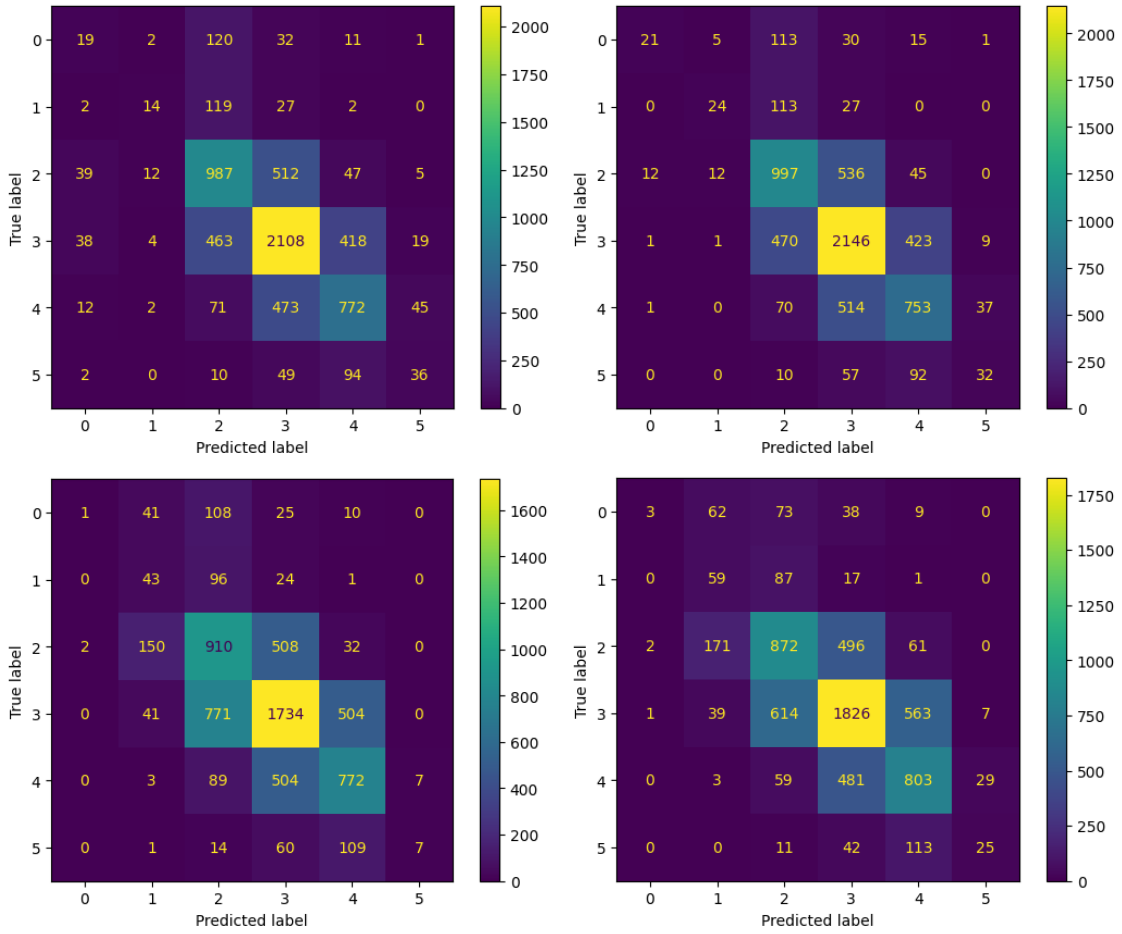

Fig. 1. LGBM with class weights and with a distinction between synthetic data (minority classes) and real data. In the first line we have LGBM(c), while in the second one LGBM(r). In the first column we have data augmentation, but in the second one we don't.

results obtained. The LGBM(c) was the best model in most of the metrics. However, its performance was lower than the previous experiments (where synthetic data participated only in training), except for the macro F1 metric for the classifiers. This results mean that the synthetic data included in the training set did not help the models in classifying correctly the original and the synthetic data in the test set.

We highlight the confusion matrices for the LGBM model in Figure 2, and we observe even greater difficulty in correctly classifying instances from classes that contain synthetic data (0, 1, and 5). In previous experiments, we noted that the synthetic data for minority classes used in training did not improve the model's performance, and we see the challenges the models face in classifying this synthetic data.

In summary, the results demonstrated that including some samples of synthetic data for the minority classes did not produce superior machine-learning models for C3. The best model was obtained using only real data with class weight assignment. Since the synthetic data generated for the minority classes was just a sample of examples, not enough to

Table 4. Results on experiments without differentiation between real and synthetic data for minority classes.

Class Weight	Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
No	LGBM(c)	**0.577106**	0.502662	0.444282	0.333626	1.617387	0.337482
No	XGB(c)	0.570086	0.497151	0.434744	0.325233	1.622525	0.328406
Yes	LGBM(c)	0.573595	**0.514047**	**0.4517**	0.352721	1.613252	0.354522
Yes	XGB(c)	**0.574616**	**0.510655**	**0.447505**	0.341503	1.618297	0.344708
No	LGBM(r)	0.438066	0.294966	0.312473	**0.376689**	**1.302303**	**0.440629**
No	XGB(r)	0.417347	0.297299	0.30502	0.356324	1.365829	0.393871
Yes	LGBM(r)	0.407157	0.284362	0.310078	**0.388374**	**1.323603**	**0.433395**
Yes	XGB(r)	0.391417	0.293079	0.297411	0.363473	1.385827	0.391286

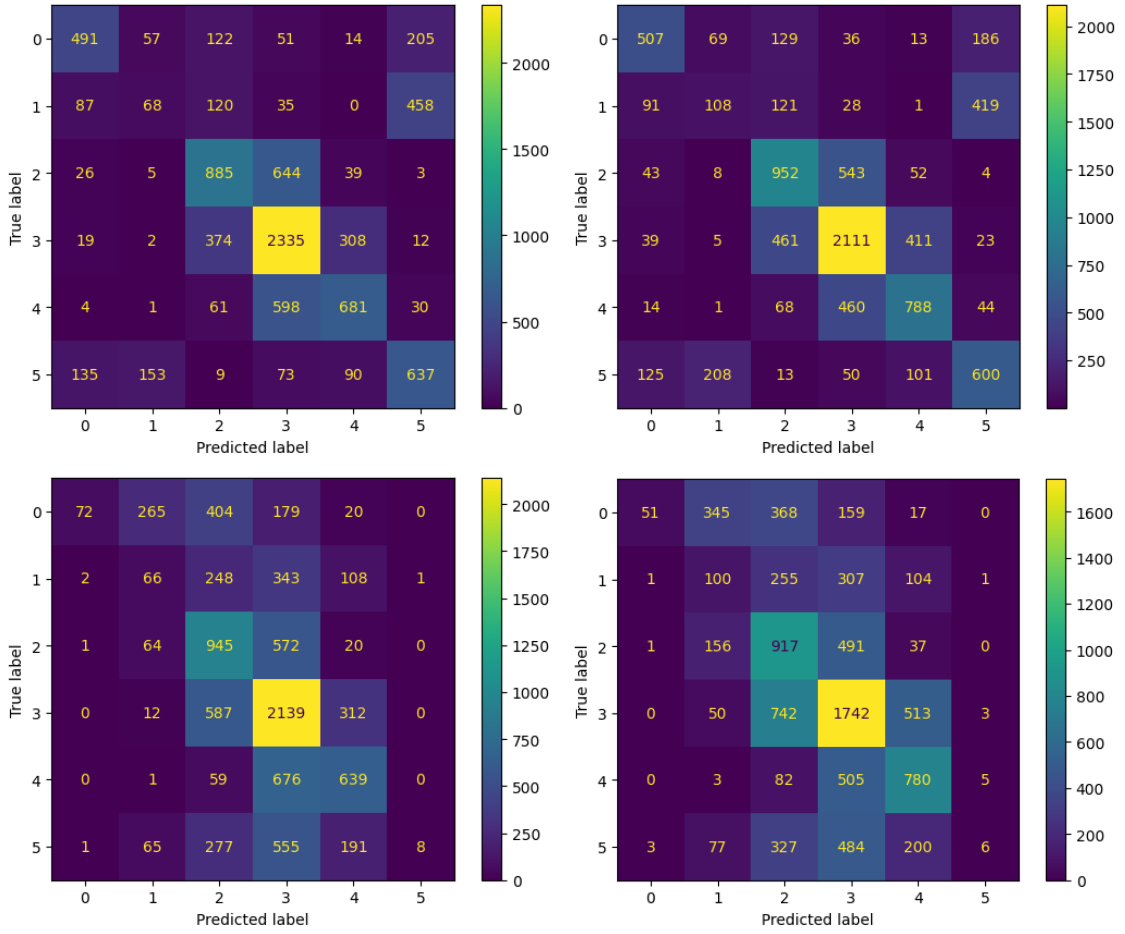

Fig. 2. LGBM without differentiation between real and synthetic data for minority classes. In the first line we have LGBM(c), while in the second one LGBM(r). In the first column we weights for class imbalance, but in the second one we don't.

balance the dataset fully, to achieve a balanced training dataset and improve model performance, we decided to explore whether a larger set of texts generated by the ChatGPT model for all classes would yield better results.

4.1.2 Fully-Balanced data augmentation. In this scenario, we used ChatGPT to produce a balanced dataset, where each class (level) has the same number of instances. To obtain this new dataset, messages were exchanged with ChatGPT using the *Openai* library, similar to the previous section, but with the following differences:

- we utilized the **gpt-3.5-turbo-0613** model, which was not available during the time of the experiments with the minority classes;
- for each of the 151 prompts available on essaybr, 16 synthetic essays were requested for each of the 6 levels (0, 40, 80, 120, 160, 200).

Consequently, a total of 14,496 essays were generated (2,416 per level). The dataset was employed in five experiments:

(1) cross-validation with real data, adding sufficient synthetic data to balance the training;
(2) cross-validation with real data, incorporating all synthetic data into the training;
(3) training on synthetic data to classify real data;
(4) training on real data to classify synthetic data;
(5) cross-validation on synthetic data.

Cross-validation on Extended EssayBr with balanced synthetic data (experiments 1 and 2). As shown in Table 1, the data for C3 is imbalanced. In the cross-validation experiments, we used five-folds, where 80% of the real data was utilized for training and the remaining 20% for testing. In this experiment, synthetic data was added at each stage of cross-validation training to ensure an equal distribution for each class. This resulted in the following distribution of synthetic data: 2292 for level 0, 2309 for 40, 1159 for 80, none for 120, 1340 for 160, and 2288 for 200.

The results achieved by the LGBM and XGB following this methodology are presented in Table 5. Even fully balancing C3 grades with synthetic data, they did not yield superior models when compared to training solely on real data, shown in Tables 2 and 3 from Section 4.1.1. The models improved only in contrast to those obtained without distinguishing between real and synthetic data for the minority classes (Section 4.1.1, Table 4).

Table 5. Synthetic data for balancing and classifying real data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	**0.595099**	**0.34539**	**0.433519**	**0.49566**	**0.862038**	**0.510311**
XGB(c)	**0.580631**	**0.343032**	**0.412061**	0.468708	0.888997	0.480967
LGBM(r)	0.54439	0.298632	0.363875	0.460282	**0.850926**	**0.497817**
XGB(r)	0.506776	0.301261	0.353866	**0.46927**	0.893439	0.481128

Table 6 shows the results when the entire balanced set of synthetic data is included in the training. As shown in Table 6, the results did not surpass those achieved by models trained without data augmentation, considering the six metrics adopted in this study. Even when the entire balanced set of synthetic data is added to the training (2416 essays for each score), there is no improvement in the performance of the models. The confusion matrices of the LGBM classifiers, which performed better in these configurations than the regressors, can be found in Figure 3. We observed that while using class weights led to better models, the addition of synthetic data (for the minority classes, to balance all classes, or 2416 for all classes) worsened performance.

Table 6. All synthetic data used in the data augmentation to classify real data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	**0.59266**	**0.342581**	**0.417903**	**0.482104**	**0.850299**	**0.506355**
XGB(c)	**0.583675**	**0.338269**	**0.406561**	**0.476704**	0.854498	**0.500739**
LGBM(r)	0.541954	0.294191	0.356069	0.455596	**0.847788**	0.497998
XGB(r)	0.502969	0.291756	0.344596	0.463607	0.89105	0.478074

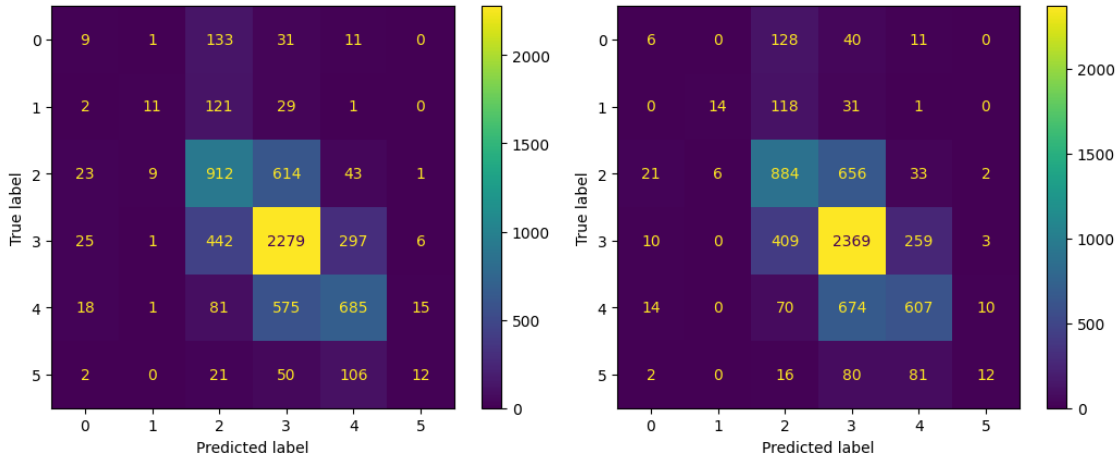

Fig. 3. Confusion matrices of the LGBM classifiers trained with real and synthetic data. The first one is from Table 5 and the second one is from Table 6.

Training on one dataset and testing on another (experiments 3 and 4). While the data imbalance in C3 poses a challenge and impacts the performance of machine learning models trained on it, it is noteworthy that using synthetic data for balancing (partially or entirely) does not lead to improvements. This raises questions about the quality of the synthetic data generated by ChatGPT.

So, we trained models with the fully-balanced synthetic data (14,496 synthetic essays - 2,416 per class) to test on real data (Table 7 and vice versa (Table 8. All six evaluation metrics significantly deteriorated, with no configuration achieving 20% accuracy. Hence, when trained on the other, a model cannot estimate one type of data (real or synthetic). This leads us to question whether real and synthetic data are too dissimilar and whether ChatGPT genuinely contributes to data augmentation. Confusion matrices for LGBM models are represented in Figure 5.

Table 7. Results achieved by models trained on the fully-balanced synthetic data to classify real data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	0.114512	0.099298	**0.092528**	0.162349	2.378908	0.323695
XGB(c)	0.103243	0.094362	0.088762	0.164217	2.34636	0.318729
LGBM(r)	**0.199482**	**0.117813**	**0.10754**	**0.237278**	**1.703551**	**0.417495**
XGB(r)	**0.162175**	**0.115403**	0.076108	**0.17764**	**1.886372**	**0.341153**

Table 8. Results achieved by models trained on real data to classify synthetic data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	0.164459	0.095659	0.011454	0.030995	**2.071667**	0.067951
XGB(c)	**0.166667**	**0.101032**	**0.013832**	**0.033357**	**2.07413**	**0.070771**
LGBM(r)	0.163631	0.09154	0.01268	**0.035985**	2.078648	**0.083814**
XGB(r)	**0.166184**	**0.1004**	**0.013241**	0.03146	2.083521	0.065456

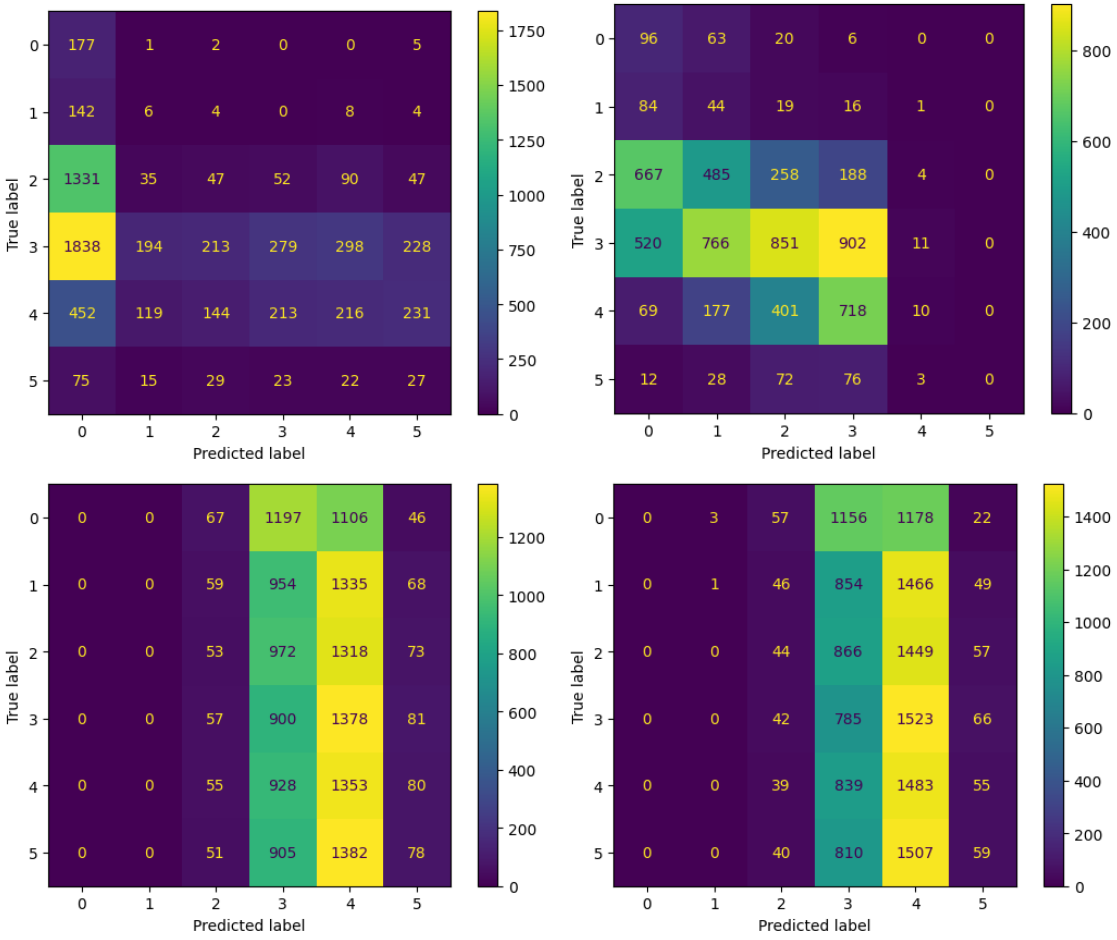

Fig. 4. Confusion matrices of the LGBM models for Table 7 in the first line, and for Table 8 in the second one. First column has classifiers and second column has regressors.

Cross-validation with synthetic data (experiment 5). The questioning about the quality of the generated essays becomes even more significant when we perform cross-validation exclusively with synthetic data (Table 9). The objective is to assess whether models trained with synthetic data can classify synthetic data in the test set. Performance remained very poor (accuracy below 20%) and was not much superior to the two previous experiments in this section (Tables 7

and 8. The result suggests poor quality of the texts generated by ChatGPT, and confusion matrices for LGBM models are in Figure 5.

Table 9. Results of the models using the cross-validation methodology with synthetic data.

Algorithm	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
LGBM(c)	**0.174737**	**0.174733**	**0.011027**	**0.013773**	1.919564	**0.013653**
XGB(c)	**0.171771**	**0.171634**	**0.011007**	**0.011353**	1.919978	**0.01125**
LGBM(r)	0.164184	0.083849	-0.000064	0.003292	**1.502001**	0.006121
XGB(r)	0.163838	0.106105	-0.000047	0.001551	**1.539597**	0.002256

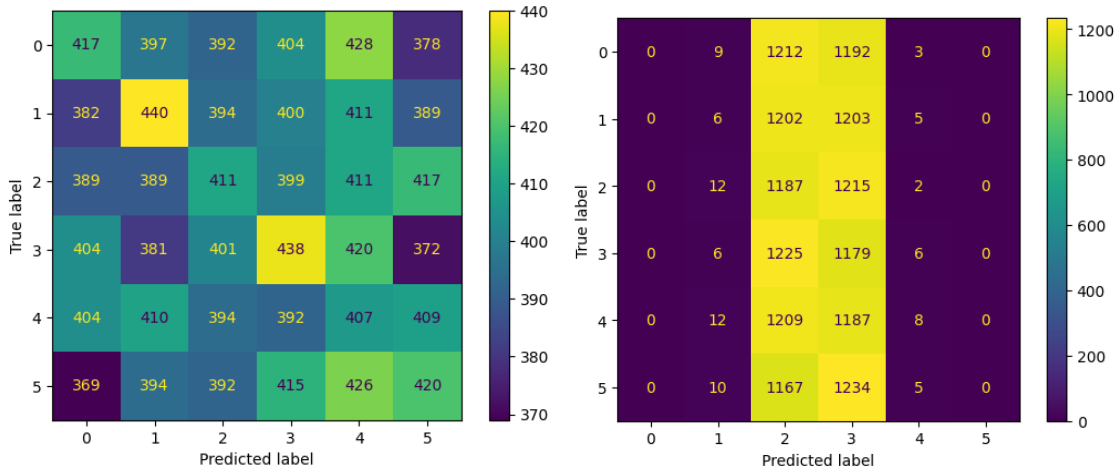

Fig. 5. LGBM models (classifier and regressor) for Table 9: using the cross-validation methodology with synthetic data.

Sections 4.1.1 and 4.1.2 introduced several approaches with data augmentation, but we observed that none of them led to an improvement in performance. The best results were achieved using only real data, with no data augmentation. Consequently, we answer RQ1: data augmentation did not improve the performance of the machine learning models for C3.

4.2 RQ2: Considering the ChatGPT as estimator, what is the best model to estimate the C3 grade?

To address RQ1, the previous section conducted experiments with machine learning models varying in different aspects: dataset, models (LGBM and XGB), classification or regression, and class weight assignment. To answer RQ2, we present the results achieved by ChatGPT and compare them with the LGBM and XGB models.

Once again, we utilized the *Openai* library with the **gpt-3.5-turbo-0613** model and the *openai.ChatCompletion.create* method with the following instructions:

- a message defining the model's role (*role: system*) with the text *Você é um avaliador de redações do ENEM do Brasil.*;
- a message with content (*role: user*) providing the essay to be assessed: *Considere a redação a seguir: TEXTO_REDAÇÃO*;

- a message with content (*role: user*) providing the essay's theme to be assessed: *Considere o texto motivador a seguir: TEXTO_MOTIVADOR*;
- a message with content (*role: user*) requesting the evalutation: *Com base no texto motivador, qual nota você daria para a redação na competência 3: 0, 40, 80, 120, 160, ou 200? Sua resposta deve ter apenas uma palavra: o inteiro representando a nota.*.

The motivator texts from the EssayBR are typically lengthy, with an average word count of 381 for the 151 themes. To reduce costs, in the analysis of EssayBR texts (6577 essays), only the title of the motivator text and the essay were provided. However, for the balanced synthetic data (14,496 essays), in order to decrease the text volume exchanged while simultaneously assessing whether this has an impact on the evaluation, the procedure was as follows:

- filter 25% of data: as each theme provides 16 essays for each score, after randomly selecting 25% of the essays for each score in each theme, we obtain a new balanced dataset with 3624 essays;
- in one scenario, only the titles of the motivator texts were passed, while in the other, the complete content was used.

So, we first requested ChatGPT to classify real data (essays from EssayBR) based on the title of the motivator text (theme). As it can be seen in the first row of Table 10, the performance was very poor, worse than all the models studied in Sections 4.1.1 and 4.1.2.

Next, we requested the correction of synthetic essays, i.e., those created by the ChatGPT itself (passing only the title of the motivator text and the essay). Surprisingly, the result was even worse, as seen in the second row of Table 10. ChatGPT had more difficulty classifying the essays it created than the real data. Finally, to check if there was any significant disadvantage in the result by passing only the title of the motivator text and essay, we decided to pass the entire motivator text and the essay. We found no significant improvement (see third row of Table 10).

Table 10. Results of ChatGPT estimating C3.

Data	Content	Accuracy	F1 macro	Kappa	QWK	RMSE	Pearson
Essaybr	Title+Essay	0.3	0.15	0.0791	0.144663	1.362132	0.16071
Synthetic	Title+Essay	0.17	0.11	0.012562	0.02379	2.193646	0.037657
Synthetic	Title+Motivator Text+Essay	0.18	0.12	0.006285	0.004858	2.242783	0.010103

Figure 6 (top) shows that ChatGPT behaves as an inefficient machine learning model when classifying real data. When ChatGPT tries to classify synthetic essays (Figures 6 (left) and 6 (right)), the performance worsens but remains comparable to models trained with real data to classify synthetics (Table 8 and Figure **??**). This suggests poor quality in the texts generated by ChatGPT. Perhaps this is why the best model obtained in this study to evaluate C3 proficiency was achieved by a model trained solely on real data.

Thus, the experiments in this section answer RQ2: ChatGPT is not capable of classifying essays for C3 better than machine learning models, and therefore, the best model found was LGBM(c) with weights trained only with real data (without data augmentation), as shown in Table 3 and Figure 1.

5 Conclusion

This article investigates strategies to enhance the estimation of the score for Competency 3 in the ENEM essay. Competency 3 is one of the five assessed by the ENEM, and its evaluation considers how the student selects, relates,

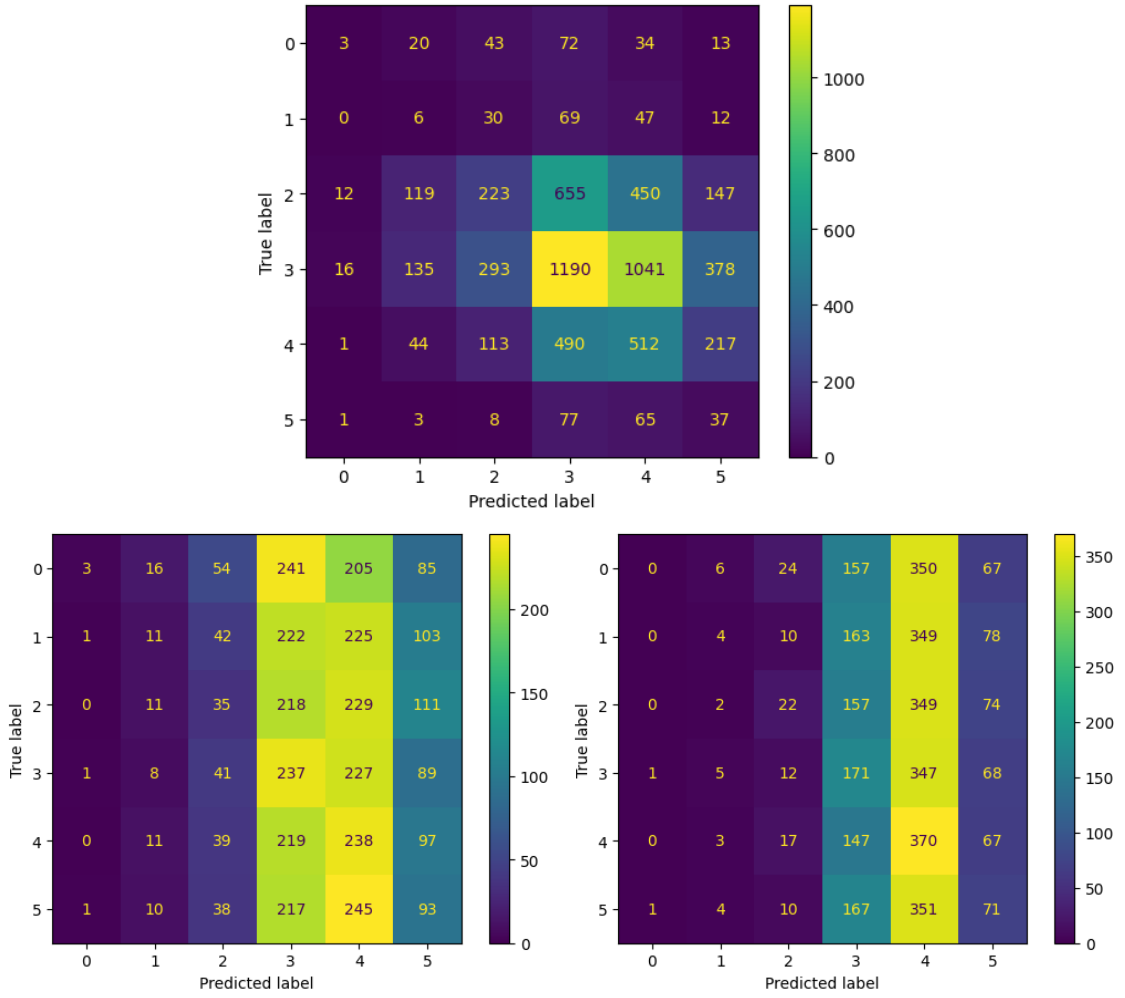

Fig. 6. Confusion matrices of the ChatGPT estimation (Table 9): considering the EssayBR essays and the prompt's title (top), considering the synthethic essays and the prompt's title (left), considering the synthethic essays and the complete prompt, with title and motivator text (right).

organizes, and interprets information, facts, and opinions in defense of a point of view. For this study, the real dataset adopted in the experiments was Essay-br, which brings an inherent difficulty in the assessment reality—data imbalance. Essays with extreme scores suffer from a lack of examples, complicating the training of machine learning models. To address this challenge, we employed LLMs to generate argumentative essays. We found that, among different LLM approaches, only ChatGPT could respond with texts that met the structure required by ENEM (argumentative essay). Thus, we created two synthetic databases: the first covered only the minority classes, while the second had a much larger volume and was entirely balanced (2416 for each score). Although several experiments were conducted with synthetic data, it was found that data augmentation did not improve the performance of machine learning models, leading us to question the quality of the texts generated by ChatGPT. Finally, we asked ChatGPT to evaluate the essays

from Essay-br and those generated by itself. The results showed that ChatGPT does not assess the texts better than machine learning models. Thus, we conclude that the best configuration was achieved by training the LGBM algorithm as a classifier using only real data and assigning weights to the classes to mitigate the imbalance.

References

[1] Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. 2022. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics* 14, 1 (apr 2022), 135–150. https://doi.org/10.1007/s13042-022-01553-3

[2] Raissa Camelo, Samuel Justino, and Rafael Ferreira Leite de Mello. 2020. Coh-Metrix PT-BR: Uma API web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*. SBC, 179–186.

[3] Flavio Carvalho, Rafael Guimarães Rodrigues, Gabriel Santos, Pedro Cruz, Lilian Ferrari, and Gustavo Paiva Guedes. 2019. Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. SBC, 24–34.

[4] Ruan Carvalho, Lucas Fernandes Lins, Luiz Rodrigues, Péricles Miranda, Hilário Oliveira, Thiago Cordeiro, Ig Ibert Bittencourt, Seiji Isotani, and Rafael Ferreira Mello. 2024. Exploring NLP and Embedding for Automatic Essay Scoring in the Portuguese. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 228–233.

[5] Keith Cochran, Clayton Cohn, Nicole Hutchins, Gautam Biswas, and Peter Hastings. 2022. Improving Automated Evaluation of Formative Assessments with Text Data Augmentation. In *Artificial Intelligence in Education*, Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Springer International Publishing, 390–401.

[6] Keith Cochran, Clayton Cohn, Jean Francois Rouet, and Peter Hastings. 2023. Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, 217–228.

[7] 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* (Online). SBC, Porto Alegre, RS, Brasil, 1403–1412. https://doi.org/10.5753/cbie.sbie.2020.1403

[8] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation. arXiv:2302.13007 [cs.CL]

[9] Semire Dikli. 2006. An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment* 5, 1 (2006).

[10] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) 2022. *A redação no Enem 2022: cartilha do participante.* Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

[11] Ruben Klein and Nilma Fontanive. 2009. Uma nova maneira de avaliar as competências escritoras na redação do ENEM. *Ensaio: Avaliação e Políticas Públicas em Educação* 17, 65 (2009), 585–598.

[12] Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2022. Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. *Journal of Information and Data Management* 13, 1 (2022), 65–76. https://doi.org/10.5753/jidm.2022.2340

[13] Jeziel C Marinho, Fábio Cordeiro, Rafael T Anchiêta, and Raimundo S Moura. 2022. Automated Essay Scoring: An approach based on ENEM competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 49–60.

[14] 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*. 509–519.

[15] Hilário Oliveira, Péricles Miranda, Seiji Isotani, Jário Santos, Thiago Cordeiro, Ig Ibert Bittencourt, and Rafael Ferreira Mello. 2022. 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*. SBC, 883–894.

[16] Husam Quteineh, Spyridon Samothrakis, and Richard Sutcliffe. 2020. Textual Data Augmentation for Efficient Active Learning on Tiny Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 7400–7410. https://doi.org/10.18653/v1/2020.emnlp-main.600

[17] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009