Comparing Feature Engineering and Deep Learning Methods for Automated Essay Scoring of Brazilian National High School Examination

Aluizio Haendchen Filho[1][a], Fernando Concatto[1][b], Hércules Antonio do Prado[2][c]
and Edilson Ferneda[2][d]

[1]Laboratory of Technological Innovation in Education (LITE), University of Vale do Itajaí (UNIVALI), Itajaí, Brazil
[2]Catholic University of Brasilia (UCB) QS 07, Lote 01, Taguatinga, Brasília, DF, Brazil

Keywords: Automated Essay Scoring, Machine Learning, Deep Learning.

Abstract: The National High School Exam (ENEM) in Brazil is a test applied annually to assess students before entering higher education. On average, over 7.5 million students participate in this test. In the same sense, large educational groups need to conduct tests for students preparing for ENEM. For correcting each essay, it is necessary at least two evaluators, which makes the process time consuming and very expensive. One alternative for substantially reducing the cost and speed up the correction of essays is to replace one human evaluator by an automated process. This paper presents a computational approach for essays correction able to replace one human evaluator. Techniques based on feature engineering and deep learning were compared, aiming to obtain the best accuracy among them. It was found that is possible to reach accuracy indexes close to 100% in the most frequent classes that comprise near 80% of the essays set.

1 INTRODUCTION

The Brazilian National High School Examination (ENEM) is an evaluation that happens annually in order to verify the knowledge of the participants about skills acquired during the high school years, including writing abilities. During the essay evaluation, two reviewers assign scores ranging from 0 to 2, in intervals of 0,5 for each of the five competencies: [C_1] Formal writing of Brazilian-Portuguese language; [C_2] Understanding the essay proposal within the structural limits of the essay-argumentative text; [C_3] Selecting, relating, organizing, and interpreting information, facts, options, and defence of a point of view; [C_4] Demonstrating knowledge of the linguistic mechanisms necessary to construct the argumentation; [C_5] Proposing of an intervention for the problem addressed based on consistent arguments.

The scoring process varies from 0 to 2 for each competence, summing 10 for the essay. A grade 0 (zero) for a competence means that the author does not demonstrate mastery over the competence in question. In contrast, a score of 2 indicates that the author demonstrates mastery over that competence. It is important to mention that two reviewers are considered in agreement when the difference between grades is less or equal than 20%.

Arguably, the essays evaluation by at least two reviewers makes the process time-consuming and expensive. According to a survey conducted by the Brazilian G1 portal, 6.1 million essays were evaluated in 2019 at a cost of US$ 4.96 per essay, reaching approximately US$ 30.27 million. This value includes the structure, logistics, and personnel needed to evaluate the national exam. On the other hand, large educational groups need to conduct training tests with students for the ENEM test. It is necessary to use at least two evaluators for each essay, which

[a] https://orcid.org/0000-0002-7998-8474
[b] https://orcid.org/0000-0003-4361-7134
[c] https://orcid.org/0000-0002-8375-0899
[d] https://orcid.org/0000-0003-4164-5828

Filho, A., Concatto, F., Antonio do Prado, H. and Ferneda, E.
Comparing Feature Engineering and Deep Learning Methods for Automated Essay Scoring of Brazilian National High School Examination.
DOI: 10.5220/0010377505750583
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 575-583
ISBN: 978-989-758-509-8

makes the process time consuming and very expensive. One of the ways to substantially reduce the cost and speed up the correction of essays is to replace one of the human evaluators by an automated process.

Automated Essay Scoring (AES) has been the subject of study for some decades. An AES system takes as input an essay and assigns a numeric score reflecting its quality, based on its content, grammar, and organization. Until recently, Machine Learning (ML) approaches using methods based on Features Engineering (FE) prevailed for predicting such outcomes (Shermis and Burstein, 2003, 2013; Dikli, 2006; Fonseca et al., 2018). Some studies have pointed out that Deep Learning (DL) AES frameworks seem to have better prediction results compared to FE-based approaches (Nguyen and Dery, 2016; Shin, 2018; Fonseca et al., 2018; Ge and Chen, 2020).

This paper presents a comparison between FE and DL results for AES, emphasizing the particular characteristics that can lead to improvements in the ENEM scoring. The comparison results point out to a solution able to automatedly score essays that, in synergy with a human evaluator, can lead to a decreasing in the set of essays that requires another human reviewer. Consequently, this approach can reduce substantially the number of required human reviewers. The solution was tested on the five ENEM competencies, however, due to space limitations, only the results with Competence C_1 are shown.

Section 2 presents some concepts of the main technologies used. Section 3 gives an overview on the related works. The details of this approach are provided in Section 4. Following, Section 5 brings a discussion on the results and the conclusion is presented in Section 6.

2 BACKGROUND

A summary of AES approaches and a description of FE-based and DL methods along with balancing techniques is presented in this section.

2.1 Automated Essay Scoring

AES is defined as the computer technology that evaluates and scores the written prose (Dikli, 2006; Shermis and Burstein, 2003). AES systems are mainly used to overcome time, cost, reliability, and generalization issues in essay assessment (Bereiter, 2003). This subject keeps attracting the attention of public schools, universities, testing companies, researchers and educators (Dikli, 2006). Usually, AES systems are dedicated to assist teachers in classroom assessment both in low and large-scale participation.

A number of studies have been driven to assess the accuracy and reliability of AES systems regarding essay assessment. Several studies have been developed in order to increase the agreement rates between AES systems and human scoring (Attali and Burstein, 2006; Foltz et al., 1999; Shermis and Burstein, 2013; Fonseca et al., 2018).

AES systems are built using several technologies and heuristics that allow for essay evaluation with fair accuracy. Moreover, unlike human evaluators, these systems maintain consistency over the assigned scores, as they are not affected by subjective factors. They can also enable faster in providing grades on essays (Shermis and Burstein, 2013).

2.2 Feature Engineering Methods

Currently, most of the research efforts for features extraction from essays are based on ML approaches (Rao and Pais, 2019). These approaches use mainly a combination of statistical and Natural Language Processing techniques to extract linguistic features. The features extracted by this method are classified with different models such as Support Vector Machines with different kernels (Shin, 2018), neural network models (Taghipour and Ng, 2016), and Gradient Boosting Trees (GBT) (Friedman, 2001; Fonseca et al., 2018).

Analysis of essays based on linguistic features is interesting not only for scoring but also for providing student feedback. Given a set of human scored essays, the features can be derived from the essays and a classifier can be trained to associate the feature values with the previously assigned score.

The statistical method Least Absolute Shrinkage and Selection Operator (LASSO), proposed by Tibshirani (1996), is an alternative to improve the accuracy and interpretability of linear regression. This is accomplished by removing the less relevant features, i.e., with lower impacts in the regression results. The gradient boosting machine proposed by Friedman (2001) works as an estimation-approximation function. It can be considered as a numerical optimization in the function space, rather than the parameter space. A connection is done among stepwise additive expansions and steepest descent minimization. A general gradient descent boosting paradigm is developed for additive expansions based on an arbitrarily chosen criterion. Specific boosting procedures are proposed for: *(i)* least-squares; *(ii)* least absolute deviation; *(iii)*

Huber-M loss functions; and *(iv)* multiclass logistic likelihood for classification. Special enhancements are derived for the particular case of regression trees, for which tools for interpretation are presented. According to Friedman (2001), the relatively high accuracy, consistent performance, and robustness of boosting may represent a noticeable advantage.

2.3 Deep Learning Methods

DL aims at solving the dependence of FE with respect to quality of the features. It is a laborious task to manually select the most informative features for such a system (Taghipour and Ng, 2016). DL aims at releasing a strong human effort in selecting features for AES.

Prediction accuracy, and interpretability of the scoring algorithms are concerns in adopting AES (Zaidi, 2016). In order to overcome such concerns, researchers have attempted to introduce improved AES frameworks (Shin, 2018). Some improvements in accuracy prediction have been obtained by means of DL algorithms or by using deep language features to ensure the model captures essay contents and the focused construct (Dong et al., 2017).

ML approaches (especially DL) for AES have shown promising prediction results (Shin, 2018; Taghipour and Ng, 2016; Dong et al., 2017). ML-based AES algorithms are heavily dependent on features selected by humans (or Feature Engineering). On the other hand, the effectiveness of DL algorithms depends only on having at least a medium or large training corpus.

Recurrent neural networks are one of the most successful DL models and have attracted the attention of researchers from many fields. Compared to feedforward neural networks, recurrent neural networks (RNN) are theoretically more powerful and are capable of learning more complex patterns from data (Taghipour and Ng, 2016). Previous studies (Kim et al., 2016; Dong et al., 2016) have demonstrated that DL AES frameworks using RNN and convolutional neural networks (CNN) can produce more robust results than the traditional models based on ML algorithms across different domains. Many algorithms have been used to demonstrate the robustness of results such as the RNN approach (Dong et al., 2017; Fonseca et al., 2018).

2.4 Classes Balancing

Class imbalance is a common problem in many application domains, including AES. The imbalance of the number of samples among the classes represents a problem for traditional classification algorithms. The problem is that these algorithms are biased by the classes' frequency distribution, which influences the prediction accuracy benefiting the more frequent classes. For example, if 25% of all essays correspond to the set of minority classes, then the algorithm will to produce a classifier with an accuracy tending to 75% (Seiffert et al., 2008).

There are different balancing techniques, like Synthetic Minority Oversampling Technique – SMOTE (Chawla et al., 2002), Adaptive Synthetic Sampling Method – ADASYN (He et al., 2008), Random Undersampling – RUS and Random Oversampling (ROS) (Yap et al., 2014). SMOTE creates synthetic examples of the minority class based on samples of this class, applying the nearest neighbour's approach. ADASYN is based on SMOTE, adding the distribution of samples on the minority class as a criterion to decide the number of synthetic examples that should be created from each sample. RUS takes the non-minority classes and randomly discard some examples in order to match the amount of the minority class. Conversely, ROS approach increases the number of non-majority classes samples by replicating them in order to match the majority class.

3 RELATED WORKS

Some studies were obtained from the literature review, considering how up to date and relevant they are to the state of art of AES in Brazilian-Portuguese language. Shin (2018) compares the effectiveness and the performance of two AES frameworks, one based on FE and the other on DL algorithms. The FE-based framework adopts support vector machines (SVM) in conjunction with Coh-Metrix features, and the second one uses the CNN approach. The results were evaluated using the Quadratic Weighted Kappa (QWK) score and compared with the results from human evaluators. CNN model outperformed the Coh-Metrix + SVM model based on the two-criterion guidelines (Writing Application and Language Convention Competencies) and produced a higher average QWK score.

Fonseca et al. (2018) pursued two directions for AES: *(i)* deep neural networks, considered the state-of-art results in the literature; and *(ii)* FE-based systems, which can benefit from domain knowledge and usually are faster to train and provide a more transparent feedback. On the FE-based method, they had trained one regressor for each competence with features extracted from the data.

The authors extracted five types of features: *(i) count metrics*: most of these features are commonplace in the literature and extract basic statistics about the text, such as number of commas, number of characters, number of paragraphs, number of sentences, sentences per paragraph ratio, average sentence length, and so on; *(ii) specific expressions*: some groups of words and expressions are expected to appear in good essays (e.g., social agents such as the *government*, *media*, *family*, *law enforcement agencies*, and *schools*); *(iii) token n-Grams*: checked in order to identify the presence of n-grams highly correlated with essay score; *(iv) POS n-Grams*: they extract a similar list of POS tag n-grams, with $2 \leq n \leq 4$, and check their presence in essays; and *(v) POS Counts*: count the occurrences of each POS tag in the text. In total, they consider a pool of 681 features values, but not all of them are relevant to each of the ENEM competencies.

For the deep neural network, they used a hierarchical neural architecture with two RNN layers. The first layer reads word vectors and generates sentence vectors, which are in turn read by the second layer to produce a single essay vector. Both recurrent layers use bidirectional Long Short-Term Memory (BiLSTM) cells. A BiLSTM is basically two LSTM, one reading the sequence from left to right and the other reading it from right to left. At each time step (each token in the first layer or each sentence in the second one), the hidden states of both LSTM are concatenated, and the resulting vector of the layer (sentence or essay vector) is obtained as the mean of all hidden states.

Taghipour and Ng (2016) developed a system called Neural Essay Assessor – NEA. It works with a recurrent neural network-based method to score the essays in an end-to-end manner. They have explored a variety of neural network models in this paper to identify the most suitable model. The best model found was a LSTM neural network trained with a regression method. The approach accepts directly an essay as input and automatically learns the features from the data.

The neural network architecture used includes five layers: *(i)* lookup table layer, that projects each word into a high-dimensional space; *(ii)* convolution layer, which extracts local features; *(iii)* recurrent layer, that works by generating embeddings (whether from the convolution layer or directly from the lookup table layer) and a representation for the given essay; *(iv)* mean over time layer, that receives the recurrent layer outputs and calculates an average vector; and *(v)* linear layer with sigmoid activation, that maps the vector generated in the mean over time layer into a scalar value. They concluded that the recurrent neural network model effectively utilizes essay content to extract the required information for scoring essays.

4 METHODOLOGY AND EXPERIMENTS

At this point, empirical results on the search for better performances are presented in terms of accuracy of FE and DL approaches in the context of AES for ENEM. It involves the following steps: *(i)* corpus and class balancing; *(ii)* FE-based approach; *(iii)* DL-based approach.

4.1 Corpus

The corpus used in the experiments was extracted by a crawling process on essays datasets from *Brasil Escola* portal (https://brasilescola.uol.com.br). Monthly, a topic is proposed in this portal and interested students submit their textual productions for evaluation. Part of the evaluated essays are then made available along with the respective corrections, scores, and comments from the reviewers. For each competence in an essay, a score between 0 and 2 is assigned, ranging in steps of 0.5.

It is also important to highlight the verification of the quality of the scores attributed by the evaluator. For this, approximately 10% of the total essays were checked by specialists in the Portuguese-Brazilian language. It was found that the agreement index between evaluators and specialists in Portuguese was close to 90%.

In order to avoid noise in the automatic classification process, the following processing steps were performed: *(i)* removal of special characters, numbers, and dates; *(ii)* transformation of all text to lowercase; *(iii)* application of morphological markers (POS tagging) using the *nlpnet* library; *(iv)* inflection of the tokens by means of stemming using the NLTK library and the RSLPS algorithm, specific for the Portuguese language; and *(v)* segmentation (tokenization) by words, sentences, and paragraphs.

Only the essays with more than fifty characters and with scores available in all competencies were considered. A set of 4,317 essays, from 2007 to 2018, was collected. The corpus has an imbalanced number of essays per grade in Competence C_1, as well as in the other four competencies, which could negatively affect the efficiency of the classifier. The first competence was chosen to illustrate how the balancing was carried out; other competencies have slightly different balances, but do not differ significantly. Fig. 1 shows the proportion of scores for each category.

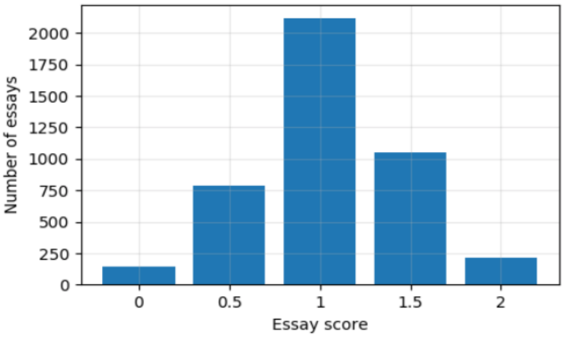


Figure 1: Class distribution in the corpus.

Better results in the experiments were achieved by using balancing techniques. For balancing, SMOTE, Adasyn, Random Oversampling, and Random Undersampling algorithms were applied. Each technique generated a new corpus that was submitted to the algorithms below.

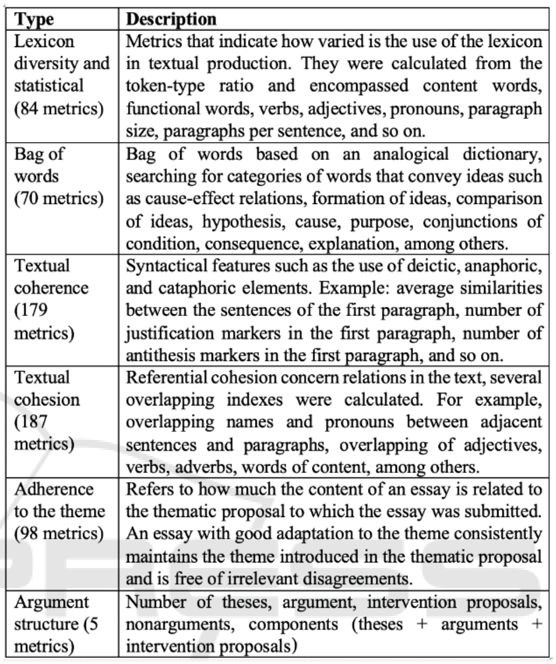
4.2 Feature Engineering Approach

This approach comprises features generation and features vector scoring.

Features Generation. It was considered 623 textual features, taking into account the results obtained by (Haendchen Filho et al., 2019), organized in five dimensions: lexical diversity, bag of words, textual cohesion, adherence to the theme, and argument structure, as shown in Figure 2. The features were submitted to a *z*-score standardization.

Features Vector Scoring. For transforming each feature vector into a grade, a function of the form F: $V \rightarrow C$ must be applied, with each $\boldsymbol{v} = (f_1, f_2, \ldots f_{623}) \in V$ representing a 623-dimensional feature vector and each $\boldsymbol{c} = (c_1, c_2, c_3, c_4, c_5) \in C$ representing a 5-dimensional vector of grades. Due to the high dimensionality of the input vector, this function must be discovered by means of inference algorithms. During the experiments, six algorithms were applied to the same problem, and concluded LASSO and GBT as the most accurate algorithms (see Figure 3). So, these were the choices for this research.

A slightly modified version of the *k*-fold cross-validation (Hastie et al., 2009) was applied. So, each cycle of the *k*-fold algorithm splits the entire training data into two disjoint subsets: a test set, containing a fraction of the full available data given by 1/*k*, and a restricted training set containing the remaining data. A stratified sampling approach was adopted, where the distribution of each class, present in the full training data, is maintained in the two subsets. This methodology guarantees the same characteristics for both the test sets and the data that will be input into the model in a deployment environment. It was used a 5-fold splitting strategy.



Type	Description
Lexicon diversity and statistical (84 metrics)	Metrics that indicate how varied is the use of the lexicon in textual production. They were calculated from the token-type ratio and encompassed content words, functional words, verbs, adjectives, pronouns, paragraph size, paragraphs per sentence, and so on.
Bag of words (70 metrics)	Bag of words based on an analogical dictionary, searching for categories of words that convey ideas such as cause-effect relations, formation of ideas, comparison of ideas, hypothesis, cause, purpose, conjunctions of condition, consequence, explanation, among others.
Textual coherence (179 metrics)	Syntactical features such as the use of deictic, anaphoric, and cataphoric elements. Example: average similarities between the sentences of the first paragraph, number of justification markers in the first paragraph, number of antithesis markers in the first paragraph, and so on.
Textual cohesion (187 metrics)	Referential cohesion concern relations in the text, several overlapping indexes were calculated. For example, overlapping names and pronouns between adjacent sentences and paragraphs, overlapping of adjectives, verbs, adverbs, words of content, among others.
Adherence to the theme (98 metrics)	Refers to how much the content of an essay is related to the thematic proposal to which the essay was submitted. An essay with good adaptation to the theme consistently maintains the theme introduced in the thematic proposal and is free of irrelevant disagreements.
Argument structure (5 metrics)	Number of theses, argument, intervention proposals, nonarguments, components (theses + arguments + intervention proposals)

Figure 2: Characteristics of the applied features groups.

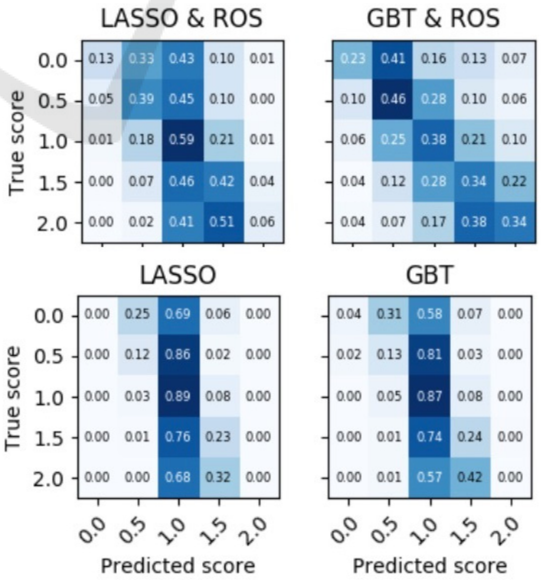


Figure 3: Correlation matrices for model pairs.

The examples in this section refer to Competence C_1 of ENEM. The distribution of occurrences in other competencies are similar to Competence C_1, except for Competence C_5, which is studied in a specific work (Haendchen Filho et al., 2019).

An analysis based on confusion matrices was carried out in order to provide an overview on the performance of the model-balancer combinations. The values of each matrix were normalized according the usual column-wise procedure, considering the extreme values as 0 or 2, and defining proportionally the remaining values. The results are shown in Fig. 3.

A few conclusions can be clearly gleaned from these results. First, when no balancing methods are used (second row of the figure), neither of the two models achieved a true positive rate of more than 4% in the classes located in the extremities (0 and 2). Thus, they cannot be applied in real situations, as they are unable to discern high- and low-quality essays from average ones. In terms of QWK, the LASSO achieved a value of 0.245, while the GBT achieved 0.285.

Second, the ROS balancing method, coupled with GBT classification model, was found to be the only kind of combination that can pinpoint low- and high-quality essays with some level of reliability. However, even with balancing, the LASSO algorithm, which represents the regressor model, concentrates the predictions in the score class 1.0, which represents the baseline. It is, therefore, inefficient in predicting scores in the extreme classes. With the balanced corpus, the LASSO achieved a QWK of 0.384, which is higher than GBT result (QWK=0.367) due to its more accurate predictions in the intermediate classes.

4.3 Deep-Learning Approach

Some researchers and developers (Dikli, 2006; Shermis and Burstein, 2013; Fonseca et al., 2018; Amorim and Veloso, 2017) share the same opinion that feature selection for AES is one of the most important tasks. According to Ge and Chen (2020), DL is a technique suitable for AES research and development and can be used to select meaningful features related to writing quality and to be applied in the AES model construction.

The results found by applying NEA framework to ENEM are here presented. The training vector was generated by a Word2vec model of the continuous bag-of-words (CBOW) variety, with 50 and 100 dimensions. The vocabulary was composed by the 4,000 most frequent words from a total of 31,953 unique words, resulting in an unknown rate of approximately 10%.

The model architecture comprises 300 LSTM as recurrent units and did not use a convolutional layer. To avoid overfitting, 50% of the outputs of the recurrent layer were dropped out; the remaining partition fed a Mean-over-Time aggregation layer (Taghipour and Ng, 2016). Finally, a fully connected layer mapped the incoming signs into a single real number – the essay's score – using a sigmoidal activation function.

The model was trained for a fixed amount of 50 epochs in each experiment. Nearly the same behaviour could be observed when using 50 or 100 dimensions in the embedding layer, with the first option offering marginally better results on average. This section concentrates on the results achieved with a 50-dimensional embedding.

Similar to the FE-based approach, a 5-fold procedure was carried out, where the data was split into five 80/20 folds before any training was proceeded. At the end of each epoch, an evaluation step was executed in which the model attempted to predict the scores for the validation set of the current fold. The results presented in this section are the best ones (in QWK) out of all 50 epochs.

The adapted NEA was trained with the corpus of 4,317 essays, without balancing. Since FE-based models assigned a score of 1 (the majority class) to the vast majority of essays, it is expected a similar behaviour using this approach. The results obtained are presented in Fig. 4, which represents epoch 23, the one with the highest QWK (0.329).

From these results, one can see that even though there is a considerable concentration on the more prevalent classes (0.5, 1.0 and 1.5), the NEA outperforms both FE-based models when no balancing is used. These results are comparable to the LASSO-ROS pair. However, these results are still lower than GBT-ROS, which produces more accurate results in the minority classes (0.0 and 2.0). These results point out that DL approaches have a high potential for scoring Portuguese written essays.

Afterwards, an experiment with the ROS balancer was executed, aiming to measure how it affects the learning process in a deep neural network. The results of the 9th epoch, which achieved a QWK of 0.336, are presented in Fig. 5.

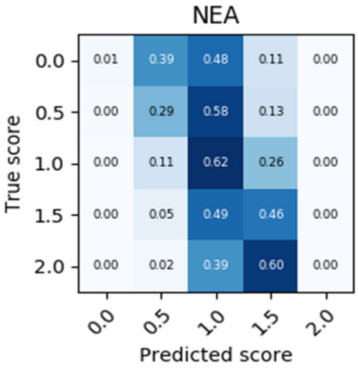


Figure 4: Confusion matrix for NEA's 23rd epoch, without class balancing.

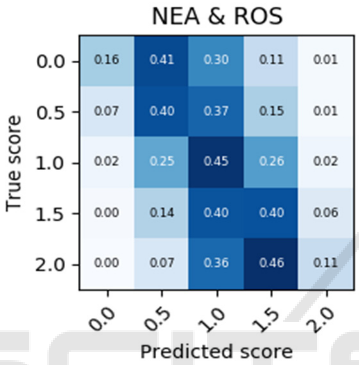


Figure 5: Confusion matrix for 9th epoch of the NEA model, with ROS.

Observing these results, it becomes noticeable that the improvement obtained by means of balancing procedure was significant, but considerably less pronounced when compared to the FE-based approach. Considering the ENEM criteria for correct predictions, an accuracy of 0.57 was achieved in both extreme classes – a value that, on average, surpasses all other models, except for GBT-ROS. Even though the DL approach produces competitive results without balancing, it has shown to be unable to surpass the accuracy of the best FE-based model in the minority classes.

5 DISCUSSION

Two main issues are discussed in this section. First, a comparison between FE and DL is presented. Next, this configuration is compared with the state of art.

5.1 On the Results

Initially, the QWK for FE and DL were calculated. FE is based on features extracted from a 623-dimensional vector of real numbers representing each essay. The resulting data set was submitted to LASSO and GBT inference algorithms. On the other hand, DL (NEA) creates its own features from the essays. Table 1 shows the computed QWK values for the three algorithms used in this work, with and without balancing.

Table 1: Computed QWK scores.

Algorithm	No balancing	With ROS
LASSO	0.245	0.384
GBT	0.285	0.367
NEA	0.329[1]	0.336[2]

[1] epoch 23
[2] epoch 9

The results demonstrate that, when the corpus is used with no balancing, the deep neural network clearly outperforms both FE-based models. As shown in the confusion matrices, the FE-based models assign a score of 1.0 (the mean) to the vast majority of essays, producing a mostly vertical figure.

NEA accuracy tends to the optimal diagonal figure, although it is still unable to precisely detect extremely good or bad essays. This accuracy improves much more when the ROS balancing procedure is applied. Both FE-based approaches exhibit a significant increase in QWK: 56.7% with LASSO and 28.8% with GBT. In the extreme classes, the largest gain was observed in the GBT method, which varied from a mean accuracy of 2% to 28.5%, while LASSO varied from 0% to 9.5%. On the other hand, DL performance did not change significantly by oversampling the corpus, showing just a small increase of 2.1% in the QWK and a very subtle change in the confusion matrix, achieving a mean accuracy of 13.5% in the extremities.

It is important to consider that, according to ENEM criteria, two reviewers are considered agreed when the difference between their grades is less or equal than 20%. It is the so-called relaxed accuracy. When the scores are in this range, the final score is considered as their mean. Table 2 presents the relaxed accuracies for each model or combination model-balancer.

Table 2: Relaxed accuracies.

Class	LASSO	LASSO-ROS	GBT	GBT-ROS	NEA	NEA-ROS
0.0	0.25	0.46	0.35	**0.64**	0.40	0.57
0.5	0.98	0.89	0.96	0.84	**100.0**	0.96
1.0	**0.99**	0.92	0.98	0.84	0.95	0.86
1.5	**0.99**	0.92	0.98	0.84	0.95	0.86
2.0	0.32	0.57	0.42	**0.72**	0.60	0.57

The relaxed accuracy in the intermediate classes, is near 100% for LASSO and for NEA, both with no balancing. The extreme classes seem to benefit from oversampling, having 0.64 and 0.72 as the best accuracies with GBT-ROS.

One can realize that, predicting better in the central classes will produce a smaller number of essays to be submitted to another evaluator, what will reduce considerably the time and costs involved.

5.2 Comparison with the State of Art

A discussion on the relation between this work and the state of art of AES for ENEM (Fonseca et al., 2018) is here presented. While Fonseca et al. (2018) reported a QWK of 0.68 for the first competence using Gradient Boosting (which achieved the best results), the approach proposed here found a QWK of only 0.384 in the best case (LASSO with oversampling). One can hypothesize that this difference is mainly a consequence of the dataset that was used by the authors: while in this work an open-access corpus of 4,317 essays was used, Fonseca et al. (2018) employed a proprietary dataset containing 56,644. Due to the small number of examples, it is likely that the models were unable to make proper generalizations from the data, therefore producing a smaller QWK value. The deep learning model used in Fonseca et al. (2018) is also proprietary, while an open source model (NEA) was applied in this work. This model was applied by Taghipour and Ng (2016) for English essays with promising results.

From the present approach, that adheres to the ENEM criteria for true positives, it is noticeable a high level of accuracy, near 100% in scores 0.5, 1.0, 1.5. In these classes, the accuracy rates are, respectively, 1.00/0.95/0.95 with DL NEA, and 0.98/0.99/0.99 with LASSO. In the extreme classes (C_1 e C_5) it was found accuracies of 0.64/0.72, respectively, by combining GBT and ROS. Notice that these classes correspond to less than 1% of the total amount of essays (see Fig. 1). Since accuracies near 100% was already reached in classes C_2, C_3 and C_4, any improvement would be residual.

6 FINAL REMARKS

Tools for helping to reduce problems related to the proficiency of the Portuguese-Brazilian language are fundamental for the development of education in Brazil. The average reading performance of Brazilian students in the exam carried out by the Program for International Student Assessment (PISA), in 2018, was below average (INEP-MEC, 2020). This deficiency is reflected in undergraduate courses, where students have difficulties to express themselves in an appropriate and logical way, which ends up compromising their learning.

In order for educational institutions to apply tests and essay writing exercises on a large scale, it is essential to reduce the costs related to correction time and, at the same time, streamline the process. In a context in which two human evaluators participate, replacing one of them by a reliable AES system is an alternative that may be feasible.

The search for an accurate system able to replace a human was one of the main objectives of this work. In a context of 5 scoring classes (0.0 / 0.5 / 1.0 / 1.5 / 2.0) for each competence, accuracies close to 100% was achieved for the three central scores, and close to 57% in the two extremity scores (0.0 / 2.0). It means that, on average, 90% of the scores assigned by the computer are correct. Another advantage is the consequent reduction in the number of essays that need to be sent to a third reviewer.

The study also showed that significant gains in accuracy can be obtained for true positives by applying balancing techniques. As class imbalance is one of the characteristics for essay grading corpora, in this work this technique has proven to be efficient for AES, and contributed to obtain required accuracies (Table 2). In addition, there are very few studies on literature exploring the corpora balancing technique in the context of AES.

Another contribution of this work is the corpus of ENEM-based essays that is made available ready to use (download from https://github.com/concatto/aes-portuguese). It is relevant for research in Portuguese, beyond the usual English. There is no equivalent corpus available for the research community.

As future works, firstly, it is suggested to combine the best aspects of each approach in an ensemble. Taking into account that some models or combinations of model-balancer techniques learn better some specific class, it is interesting to build a model with these combinations and take advantage from the particular accuracies. Other interesting future work is to improve the predictive quality of the features. Although the high level of relaxed accuracy - near 100% - had been reached for the dominant classes, there is still room to improve the QWK scores in the extremes. Finally, the accuracy on ENEM results could be enabled by taking a bigger corpus and exploring new features.

REFERENCES

Amorim, E. C. F., Veloso, A., 2017. A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. *15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 94-102. DOI: 10.18653/v1/E17-4010.

Attali Y., Burstein, J., 2006. Automated Essay Scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Bereiter, C., 2003. Foreword. In Mark D. Shermis and Jill Burstein (Eds.), *Automated essay scoring: a cross disciplinary perspective*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. vii-x.

Chawla, N. V., Bowyer, K. W., Hall, Lawrence, O. W., Kegelmeyer, P., 2002. SMOTE: Synthetic Minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321-357. DOI: 10.1613/jair.953.

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

Dong, F., Zhang, Y., Yang, J. 2017. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. *21st Conference on Computational Natural Language Learning – CoNLL 2017*, pp. 153-162. DOI: 10.18653/v1/K17-1017.

Foltz, P. W., Laham, D., Landauer, T. K., 1999. Automated Essay Scoring: Applications to educational technology. *World Conference on Educational Multimedia, Hypermedia and Telecommunications – EdMedia'99*, pp. 939-944.

Fonseca, E. R., Medeiros, I., Kamikawachi, D., Bokan, A., 2018. Automatically grading Brazilian student essays. *13th International Conference on Computational Processing of the Portuguese Language – PROPOR 2018*, pp. 170-179. DOI: 10.1007/978-3-319-99722-3_18.

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189-1232.

Ge, S., Chen, X., 2020. The Application of Deep Learning in Automated Essay Evaluation. In *Emerging Technologies for Education*. LNCS 11984, Springer, pp.310-318. DOI: 10.1007/978-3-030-38778-5_34

Haendchen Filho, A., Concatto, F., Nau, J., do Prado, H. A., Imhof, D. O., Ferneda, E., 2019. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. *Procedia Computer Science*, 159:764-773.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., Aluísio, S. M., 2017. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *Symposium in Information and Human Language Technology – STIL 2017*, pp. 122-131.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2nd Edition.

He, H., Bai, Y., Garcia, E. A., Li, S., 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322-1328. DOI: 10.1109/IJCNN.2008.4633969.

INEP-MEC, 2018. *Relatório Brasil no Pisa 2018*. Diretoria de Avaliação da Educação Básica. Ministério da Educação. Brasilia. Brazil. http://portal.inep.gov.br/documents/186968/484421/RELAT%C3%93RIO+BRASIL+NO+PISA+2018/3e89677c-221c-4fa4-91ee-08a48818604c?version=1.0

Kim, Y. Jernite, Y., Sontag, D., Rush, A. M., 2016. Character-aware neural language models. *Thirtieth AAAI Conference on Artificial Intelligence – AAAI-16*, pp. 2741-2749.

Nguyen H., Dery, L., 2016. *Neural Networks for Automated Essay Grading*. CS224d Stanford Reports. https://cs224d.stanford.edu/reports/huyenn.pdf.

Rao, R. S., Pais, A. R., 2019. Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Comput & Applications* 31:3851-3873. DOI: 10.1007/s00521-017-3305-0.

Seiffert, C., Khoshgoftaar, T. M., van Hulse, J., Napolitano, A., 2008. Building Useful Models from Imbalanced Data with Sampling and Boosting. *Twenty-First International FLAIRS Conference*, pp. 306-311.

Shermis, M. D., Burstein J. (Eds.), 2003. *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Shermis, M. D., Burstein J., 2013. *Handbook of Automated Essay Evaluation: Current applications and new directions*. Routledge, New York, NY.

Shin, E., 2018. *A Neural Network approach to Automated Essay Scoring: A Comparison with the Method of Integrating Deep Language Features using Coh-Metrix*. Master Thesis. Department of Educational Psychology University of Alberta, Canada. DOI: 10.7939/R3V11W25D.

Taghipour K., Ng, H. T., 2016. A Neural Approach to Automated Essay Scoring. *2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882-1891.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, 58(1):267-288.

Yap, B. W., K. A. Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., Abdullah, N. N., 2014. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *First International Conference on Advanced Data and Information Engineering*, pp. 13-22. Lecture Notes in Electrical Engineering 285, Springer. DOI: 10.1007/978-981-4585-18-7_2.

Zaidi, A. H., 2016. *Neural Sequence Modelling for Automated Essay Scoring. Master Thesis*. University of Cambridge. https://www.cl.cam.ac.uk/~ahz22/docs/mphil-thesis.pdf.

Zou, Will Y., Socher, R., Cer, D., Manning, C. D., 2013. Bilingual word embeddings for phrase-based machine translation. IN *Conference on Empirical Methods in Natural Language Processing*, pp. 1393-1398.