Comparative Study of Results Obtained with Machine Learning Algorithms in Automatic Correction of Essays

Fernando Concatto[1], Jonathan Nau[1], Marilu Leidiane Pazzetto[2], Aluizio Haendchen Filho[1,2]

[1]Laboratório de Inteligência Aplicada – UNIVALI
Caixa Postal 360 – CEP 88302-202 – Itajaí – SC – Brasil

[2]Núcleo de Inteligência Artificial e Sistemas Inteligentes – UNIFEBE
Brusque – SC – Brasil

{fernandoconcatto,aluizio.h.filho}@gmail.com

Abstract. *Every year, high school students in Brazil take the National High School Exam (ENEM), which is comprised of a series of multiple-choice questions and a handwritten essay. To assign a grade to the essay, at least two reviewers must evaluate it individually; thus, the grading process is extremely time-consuming and represents a large cost to the government. Therefore, in this paper, we sought to identify viable methods to automatically assign a grade to ENEM essays employing machine learning algorithms, considering a corpus extracted from an online practicing platform and a 618-dimensional hand-engineered feature vector. We also investigate the impact of feature selection methods in the general performance of the grading system.*

Resumo. *Todo ano, estudantes no Brasil realizam o Exame Nacional do Ensino Médio (ENEM), composto por uma série de questões de múltipla escolha e uma redação. Para atribuir uma nota à redação, pelo menos dois revisores devem avaliá-la individualmente; portanto, o processo de correção consome muito tempo e representa um grande custo ao governo. Portanto, neste artigo, buscou-se identificar métodos viáveis para corrigir automaticamente redações do ENEM através de algoritmos de aprendizagem de máquina, considerando um corpus de texto extraído de uma plataforma online e um vetor de atributos de 618 dimensões construído à mão. Também investigou-se o impacto de métodos de seleção de atributos no desempenho do sistema de avaliação.*

1. Introduction

The national high school examination (known as ENEM) is an evaluation that happens annually in Brazil in order to verify the knowledge of the participants about various skills acquired during the school years. There are four exams consisting of multiple-choice tests, encompassing diverse contents, and a manuscript essay. The multiple-choice or objective questions are evaluated according to the response indicated, but the essay needs to be evaluated by at least two reviewers, which makes the process time-consuming and expensive.

During the essay evaluation, two reviewers assign scores ranging from 0 to 200, in intervals of 40, for each of the five competencies that make up the evaluation model. Score 0 (zero) indicates that the author of the text does not demonstrate mastery over the competence in question. In contrast, score 200 indicates that the author demonstrates mastery over competence. The competencies evaluated are:

1. Domain of the standard norm of the Portuguese language;
2. Understanding the essay proposal;
3. Organization of information and analysis of text coherence;
4. Demonstration of knowledge of the language necessary for argumentation;
5. Elaboration of a proposed solution to the problems addressed, respecting human rights and considering the socio-cultural diversities.

Systems for automatic grading of essays are built using several technologies and heuristics that allow evaluating with certain accuracy the quality of essays. Moreover, unlike human evaluators, these systems maintain consistency over the assigned scores, as they are not affected by subjective factors. They also help to reduce costs and enable faster feedback to the student-practicing essay [Nau et al. 2017]. Therefore, this paper aims at exploring a variety of algorithms and preprocessing methodologies, with the ultimate goal of implementing a computational system able to properly grade ENEM essays, minimizing the need for human intervention.

2. Background

The automated essay evaluation is a multidisciplinary field that has researches in: (i) cognitive psychology; (ii) computer science; (iii) educational measurement; (iv) linguistics; and (v) written research [Shermis and Burstein 2013]. Currently, mainly for English, there is already a lot of research in the field and some commercial writing evaluation applications, among them, according to Dikli (2016), Intelligent EssayTM, CriterionSM, E-rater®, IntelliMetricTM, MY Assessor!®.

However, in the Brazilian scenario, the automated essay evaluation changes completely, for the evaluation of essays in Portuguese, it does not have commercial software and there is little research in this field. The only work that evaluates the writing in all the competences of the ENEM are: (i) Amorim and Veloso (2017); and (ii) Fonseca, Medeiros, Kamikawachi and Alessandro Bokan (2018). If expand the search for research that evaluate a single competency, you will find: (i) off-topic essay detection [Passero 2017]; (ii) spelling and grammatical errors [Júnior; Spalenza; Oliveira 2017]; and (iii) textual coherence and cohesion [Lima et al. 2018].

In the paper by Fonseca et al. (2018), the authors presented two approaches for automated essay scoring in Portuguese, the first is the application of a deep neural network and the second the method based on features. The neural network has two recurrent layers and was trained for only two epochs, with batches of 8 essay. In the features approach the authors used the gradient boosting and linear regression algorithms, with statistical features, POS tagging counts and specific expressions. In the first four competitions both approaches reach similar levels of performance, but in the fifth competition the deep neural network obtained better performance.

3. Methodological Procedures

The research work was developed by means of the following steps: *(i)* organization of the corpus; *(ii)* extraction and normalization of features; *(iii)* training algorithms and *(iv)* validation. These steps are described as follows.

3.1. The Corpus of Essays

The essays used to construct the corpus that enabled our experiments were obtained through a crawling process of essays datasets from the UOL and Brazil School portal [UOL Educação 2018, Brasil Escola 2018]. Both portals have similar processes for the accumulation of essays: monthly a theme is proposed and interested students submit their textual productions for evaluation. Part of the essays evaluated are then made available on the portal along with the respective corrections, scores and comments of the reviewers. For each essay, a score between 0 and 2 is assigned, varying in steps of 0.5 for the 5 competences corresponding to the ENEM evaluation model.

In order to avoid noise in the automatic classification process, we perform the following processing steps:

1. Removal of special characters, numbers and dates;
2. Transformation of all text to lowercase;
3. Application of morphological markers (POS tagging) using the nlpnet library;
4. Inflection of the tokens by means of stemming using the NLTK library and the RSLPS algorithm, specific for the Portuguese language;
5. Segmentation (tokenization) by words, sentences and paragraphs;

In addition to these steps, only the essays with more than fifty characters and whose scores available in all competencies were considered. Table 1 presents the general characteristics of the corpus after preprocessing.

Table 1. General metrics on the essays corpus

Metric	Value
Nº of essays	2164
Nº average words per essay	269.84

The scores attributed by the UOL site for each of the competences were reviewed by two Portuguese teachers, who examined 400 of the total number of essays in the corpus. In some cases, there were discrepancies between the scores awarded by the UOL site and by the reviewers; when both reviewers pointed out the same new grades, the scores were corrected. This occurred in approximately 20% of the grades assigned in each competence, and the highest percentages occurred in competences 1 and 5. The corrected scores were considered in one of the hypotheses tested.

3.2. Feature extraction, normalization and selection

Similarly to Júnior, Spalenza and Oliveira (2017), each essay was represented as a feature vector. In total, 618 metrics were considered. The features comprise several dimensions of the five competences of ENEM. lexical diversity, readability indexes, bag of words, counting of connectives and measures of word overlap between sentences and

between paragraphs, among others. In the bag of words group, some new bags related to the dissertative argumentative quality of the text were incorporated. This improvement attempt was made based on analogical dictionary of the Portuguese language [Azevedo 1999]. The new features generated were considered in the test of hypotheses.

3.2.1. Feature selection

Given the large number of variables extracted from the essays and also due the importance of the ability to comprehend the automatic grading process, the study presented in this paper considered an approach to reduce the number of features utilized in the inference step, aiming to improve the interpretability of the model while keeping its accuracy. Guyon and Elisseeff (2003) also suggest that selecting a subset of input features might reduce the training time of the model, diminish storage requirements and possibily increase its predictive performance.

The strategy analyzed in this work involves a variable ranking method with scores computed through the ordinary least squares algorithm. First, every input feature was standardized such that its mean was equal to zero and its standard deviation was set to one across all training samples. This step ensures that no variable receives a higher or lower score merely due to their scale (e.g. having only large positive values while others have smaller positive and negative values). The transformation was implemented with the Z-score, given by:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x represents a single value of the feature, μ indicates the mean of all values of the feature and σ represents its standard deviation. Afterwards, the ordinary least squares algorithm was applied using the entire dataset, producing a vector of coefficients, containing one value for each feature (as discussed in Section 3.3.1). The absolute value of the coefficient, averaged across the five grading criteria, was then defined as the importance score of the respective feature.

3.3. Inference algorithms

To transform each of the feature vectors into a grade, a function of the form $F : V \rightarrow C$ must be applied, with each $\mathbf{v} = (f_1, f_2, \ldots, f_{618}) \in V$ representing a 618-dimensional feature vector and each $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5) \in C$ representing a 5-dimensional vector of grades. However, due to the high dimensionality of the input vector, such a function must be discovered through inference algorithms. This goal is typically accomplished through regression analysis, whereby a relationship between a set of independent variables and one dependent variable is automatically computed. However, since only a single dependent variable may be specified in this approach, the problem was split into five sub-problems, such that each competence is associated with its own function, in the form $F_i : V \rightarrow \mathbb{R}$; therefore, in total, five different functions must be estimated.

A wide array of regression methods are available in the literature [Efron and Hastie 2016]. To identify the most appropriate algorithm for computing the five grading functions, four methods were studied: LASSO, Linear Regression, Support Vector Machines and Regression Trees. The following sections describe each of the algorithms.

3.3.1. Linear Regression

Linear Regression, in its simplest form, involves fitting a line to an *n*-sized set of (x, y) points, producing a function of the form $f(x) = \beta_0 + \beta_1 x$. The fitting process is typically done by the least squares algorithm, which was developed by Gauss and Legendre circa 1800, and works by finding the values of β_0 and β_1 that minimize the sum of squared deviations over all *n* points [Efron and Hastie 2016].

The Linear Regression algorithm can also accept a *k*-dimensional vector as input instead of a single scalar value; when this is the case, the algorithm is called Multiple Linear Regression, and the function takes the form $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. Since the set of features extracted from the essays possesses 618 dimensions, this variation of the algorithm was adopted in this study. This method was chosen due to its simplicity, thus forming a baseline for other, more sophisticated algorithms.

3.3.2. LASSO

The LASSO was proposed by Tibshirani (1996) as a means to improve the interpretability and the accuracy of the Linear Regression method by reducing the number of variables used in the model. This is accomplished by introducing a constraint in the utility function being minimized, which limits the sum of the absolute values of the coefficients to a parameter called *t*. Formally, this constraint may be represented as $|\beta_1| + |\beta_2| + \ldots + |\beta_k| \leq t$, and it incurs in a general reduction in the values of each coefficient, setting some of them to 0 and thus attaining a form of variable selection and enabling a comparison with other methods of feature selection.

Equivalently, the constraint can be represented as a penalty term ψ added to the original function, defined as $\psi = \alpha (|\beta_1| + |\beta_2| + \ldots + |\beta_k|)$, with $\alpha > 0$. Higher values of α cause more coefficients to be nullified, thus simplifying the model, while $\alpha = 0$ reduces the method to the original least squares algorithm. For the purposes of the study presented in this paper, defining $\alpha < 0.1$ generally produced better results.

3.3.3. Support Vector Machines

Support Vector Machines were introduced by Cortes and Vapnik (1995) as a method for solving binary classification problems with a learning algorithm. Fundamentally, the SVM approach involves mapping the input space into a high-dimensional feature space, making the task of separating the training data into two classes easier, and finding the hyperplane that produces the largest possible margin between the closest data points of both classes; thus, the algorithm is typically defined as an optimization problem.

Drucker et al. (1996) adapted the original SVM approach such that it could be applied to regression analysis. In this formulation, the goal is to minimize the L2 norm of the coefficients while keeping the error of each data point below a new parameter, called ε. However, this thresholding constraint can be too restrictive; therefore, slack variables are typically added to the objective function problem, making it less restrictive while maintaining the optimality-seeking principles. This method was selected due to its capacity to perform nonlinear classification effectively.

3.3.4. Regression Trees

Decision trees work by constructing a model that possesses a number of nodes, each associated with an input variable and a condition that specifies which path towards another node should be chosen, and a set of leaves, which are characterized as a special type of node that define to which class the current sample belongs (classification trees) or the value of the output variable (regression trees).

Various algorithms have been developed to construct the best possible tree structure automatically, such as ID3 and C4.5. However, some of them support only classification trees; therefore, in this work, the CART algorithm was selected due to its capacity to learn regression trees [Breiman et al. 1984].

4. Results and Analysis

The main objectives of this study were to: *i)* measure the impact of grade corrections, discussed in Section 3.1, and the revised set of bag-of-words features, presented in Section 3.2; *ii)* analyze the performance of the proposed approach for selecting a subset of the 618 features, described in Section 3.2.1; and *iii)* compare the four inference algorithms discussed in Section 3.3 considering a number of different scenarios. Of special interest is the comparison of the results obtained by our methodology across all five evaluation criteria of the ENEM, due to the highly differing textual characteristics that each individual criterion must consider.

4.1. First experiment: manual fine-tuning

The first major group of experiments involved applying the inference algorithms over the essays with and without the corrected grades, as well as with the base and revised bags-of-words. Its results are displayed in Table 1, where each cell contains the Mean Absolute Error (MAE) of the prediction, calculated by taking the average of the absolute values of the difference between the predicted and the actual grade of each essay. For this experiment, the k-fold cross validation method was utilized, with k = 10. Implementations of the algorithms were provided by the scikit-learn library for the Python programming language [Pedregosa et al. 2011]. For the LASSO, a value of 0.01 was used for the parameter α; for the SVM, the kernel Radial Basis Function (RBF) was utilized and ε was set to 0.1.

The observed results provide a set of clear implications. First, both the LASSO and SVMs present lower values of the error metric when compared to the other two algorithms. This can be explained by the fact that both approaches are known to handle high dimensional input spaces in an efficient manner. The LASSO offers a slightly lower error than SVMs, but the difference is relatively small (always less than 0.01).

Another clear result is that using the corrected grades instead of the original ones whenever available offers a small but consistent improvement in the error rates, except when using regression trees and the base bags-of-words; in this configuration, the error increases in all but the fifth criterion. For the two types of bags-of-words, an irregular behavior can be observed: in the LASSO, which discards irrelevant features, the difference was negligibly small, indicating that both sets of features were discarded;

when using linear regression, improvements were observed in all cases, but for SVMs the error increased in every test; and for regression trees, the results were mixed.

Table 1. MAE values for each combination of grade correction (Original/Corrected) and groups of bags-of-words (Base/Revised).

C	Selector	Lasso		Lin. Regression		SVM		Regression Tree	
		Base	Revised	Base	Revised	Base	Revised	Base	Revised
1	Original	0.3399	0.3397	0.3831	0.3687	0.3479	0.3544	0.4801	0.4891
	Corrected	0.3360	**0.3357**	0.3749	0.3604	0.3429	0.3485	0.4855	0.4690
2	Original	0.3608	0.3621	0.4023	0.3908	0.3663	0.3711	0.4949	0.5100
	Corrected	**0.3560**	0.3574	0.3982	0.3866	0.3621	0.3668	0.5030	0.5097
3	Original	0.3712	0.3710	0.4073	0.3975	0.3695	0.3789	0.4961	0.4956
	Corrected	0.3634	0.3632	0.3968	0.3879	**0.3605**	0.3699	0.4942	0.4864
4	Original	0.3936	0.3939	0.4271	0.4223	0.3956	0.4081	0.5428	0.5284
	Corrected	**0.3873**	0.3875	0.4193	0.4153	0.3887	0.4007	0.5464	0.5261
5	Original	0.4219	0.4219	0.4623	0.4465	0.4211	0.4225	0.5934	0.5806
	Corrected	0.4151	0.4152	0.4547	0.4378	**0.4141**	0.4145	0.5566	0.5846

Finally, the five ENEM grading criteria have an increasing level of difficulty in the context of regression analysis, as the error for criterion i was always lower than the error for criterion $i + 1$ in all of the four algorithms. Criteria 4 and 5 have a particularly higher error rate, indicating that they might have a more accentuated level of subjectivity when compared to the other three.

4.2. Second experiment: feature subset assessment

The next experiments analyzed the influence of using less features in the accuracy of the inference algorithms. The main goal of this study was to identify how each algorithm responds to a reduction in the number of features, as well as determining whether or not there exists an optimal set of features, satistying both the requirement of providing a low mean absolute error and maintaining a fair level of explicability in the results.

To this end, all 618 features were ranked in accordance to the method described in section 3.3.1. Then, the two features with the highest ranking were selected, and all four algorithms were applied to the full dataset (in both its original and corrected versions), maintaining a 10-fold cross validation strategy utilized in the previous experiment. Then, the next two leading features were added to the input vector, and the inference process was executed again. This procedure was repeated until the input space reached 500 dimensions. The results obtained are displayed in Figure 1, where each grading criterion is plotted individually. Tests using the regression tree algorithm were omitted due to its high error rates, which were causing too much clutter in the charts. It also did not display much change in error as features were added; instead it fluctuated around the same values observed in Table 1 for all subsets.

The results obtained show that using less features actually reduces the error rates significantly, especially in the case of the linear regression algorithm. Both the LASSO and SVMs display a tendency towards stabilization after the input vector reaches about 200 dimensions; this behavior is explained by the inherent ability of both algorithms to ignore irrelevant features, although a slightly increased error rate can be observed when the input vector attains its maximum dimensionality. A decrease in the performance of the LR algorithm can be noticed after approximately 300 dimensions in all criteria, presumably due to a high amount of redundant features. Finally, this experiment also confirmed the superiority of the corrected version of the dataset in every scenario.

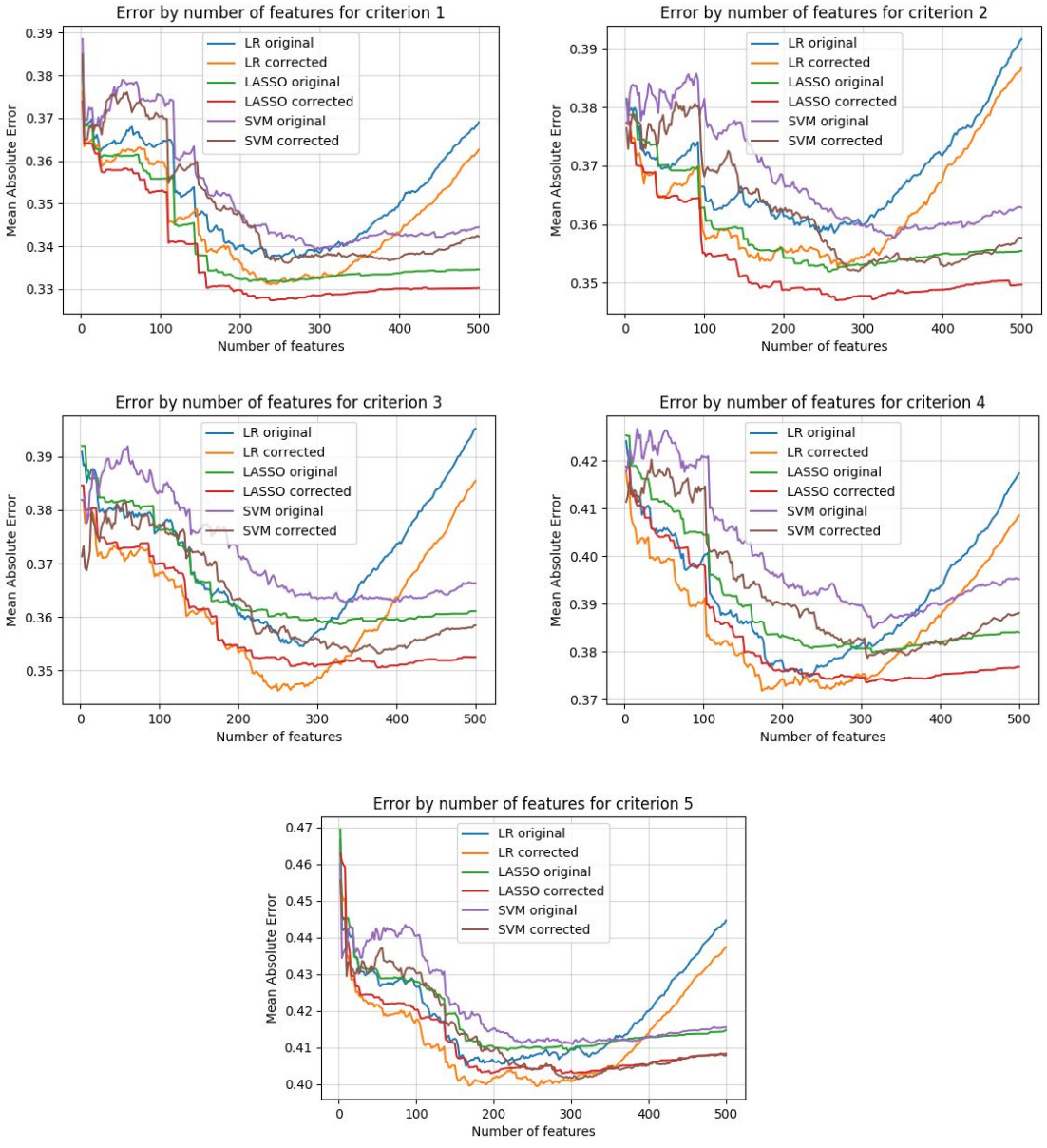

Figure 1. Relationship between the number of selected features and the mean absolute error observed over all five criteria.

5. Concluding Remarks

In this paper, a comparative study between four inference algorithms applied to the problem of automatically grading argumentative essays with respect to the five evaluation criteria of the Brazilian ENEM has been presented. Experimental evidence reveals that both the LASSO algorithm and SVMs produce a lower error than the alternatives tested in every criterion when the full set of features extracted from the text is considered. However, when a feature ranking method based on the absolute values of the coefficients estimated from the standardized features by the ordinary least squares algorithm was applied, models fit through linear regression were shown to have a lower mean absolute error than the other two algorithms in three of the five criteria when using only the 200 to 300 features that received the highest ranking. Additionally, considering that the dataset utilized was extracted from an unofficial online source, a second evaluation by human experts has been shown to provide a reduction in the mean prediction error.

In average, the best results obtained show that the methodological process presented in this study could be applied in practical, real-world contexts, since error values stood below the maximum allowed disparity between human graders (0.5 points) by the ENEM regulation; however, when individual essays are considered, the difference between predicted and correct grades might exceed the threshold of 0.5 points, especially for essays whose expected grades lie in the extremities of the distribution, i.e. between 0 and 2 points, due to the a disproportionately elevated number of essays with a grade of 1.

Future research directions include analyzing other alternatives to perform the feature selection process, such as metaheuristics and other ranking functions (e.g. entropy, mutual information); testing other regression methods such as artificial neural networks; and investigating strategies to reduce the impact of an unbalanced dataset in the training step.

References

Amorim, E. C. F e 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, pages 94–102, Valencia, Spain.

Azevedo S.S.F. (1999). Dicionário analógico da Lingua Portuguesa. In Portuguese, Lexikon Editora, São Paulo (1999).

Brasil Escola (2018). “Essay Data”.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Dikli, S. (2006). An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment, 5(1).

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Filho, A. H., do Prado, H. A., Ferneda, E., and Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. Procedia Computer Science. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 22nd International Conference, KES, Belgrade, Serbia.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonc¸alo Oliveira, H., and Paetzold, G. H., editors, Computational Processing of the Portuguese Language, pages 170–179, Cham. Springer International Publishing.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3 (Mar):1157–1182.

Júnior, C. R. C. A., Spalenza, M. A., and Oliveira, E. (2017). Proposta de um Sistema de Avaliação Automática de Redações do ENEM Utilizando Técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural. In VIII Computer on the Beach, Florianópolis.

Nau J, Haendchen Filho A., Passero G. (2017). Uma ferramenta para identificar desvios de linguagem na língua portuguesa. *Proceedings of Symposium in Information and Human Language Technology.*

Passero, G. (2017). Detecção de fuga ao tema em redações na língua portuguesa. Dissertation (Master in Applied Computation) - University of Vale do Itajaí - UNIVALI, Itajaí.

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

Shermis, M. D. and Burstein, J. (2013). Handbook of Automated Essay Evaluation: Current Applications and New Directions. Routledge, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), pages 267–288.

UOL Educação (2018). “Essay Data”.