

# A Bayesian Classifier to Automatic Correction of Portuguese Essays

Bruno Smarsaro Bazelato  
UFES - CEUNES  
Rodovia BR 101 Norte, Km. 60  
São Mateus, Espírito Santo  
+552733121584  
brunosmarsaro@gmail.com

Evelin C. F. de Amorim  
UFES - CEUNES  
Rodovia BR 101 Norte, Km. 60  
São Mateus, Espírito Santo  
+552733121584  
evelin.amorim@ufes.br

## ABSTRACT

The essay correction has been a big challenge for universities, government and professors, because of the high cost in the assessment of essays. For instance, the Brazilian National Exam for High School, also known as ENEM (Exame Nacional do Ensino Médio), is applied in order to evaluate the students after High School. ENEM evaluate the writing skill through a general topic essay. So as to evaluate all the ENEM essays, the last ENEM edition hired 40% more human evaluators than in previous edition. One way to diminish the cost of human evaluators is perform an automatic essay correction. The Automatic Essay System (AES) has been evolved since 1990s, but most of them assess English essays. The approach proposed in this paper describes a Bayesian Classifier that assesses Portuguese essays. The two main contributions of our approach are a public database of Portuguese essays, which grades vary from 0 to 10; and a baseline classifier for Portuguese Essays.

## RESUMO

A correção de redações tem sido um grande desafio para as universidades, governo e professores, devido ao alto custo na avaliação das redações. Por exemplo, o Exame Nacional do Ensino Médio (ENEM) é aplicado para avaliar os estudantes após o ensino médio. O ENEM avalia a habilidade de escrita através de um tema geral para as redações. A fim de avaliar todas as redações do ENEM em sua última edição foram contratados 40% a mais de corretores do que em sua edição anterior. Uma forma de diminuir o custo com os corretores seria um sistema de correção automático de redações. O Sistema Automático de Redações (AES) tem se evoluindo desde os anos 90, porém grande parte destes sistemas avaliam redações na Língua Inglesa. A abordagem proposta neste artigo descreve um Classificador Bayesiano que avalia redações na Língua Portuguesa. As duas principais contribuições para nossa

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*Conference '10*, Month 1–2, 2010, City, State, Country.  
Copyright 2010 ACM 1-58113-000-0/00/0010 ...\$15.00.

abordagem são um corpus de redações em Português, nas quais suas notas variam de 0 a 10; e um classificador de referência para redações em Português.

## Categories and Subject Descriptors

I.7.5 [Document Text Processing]: Document Capture – document analysis.

## General Terms

Algorithms, Performance.

## Keywords

Automatic Essay System; Bayesian Classifier; Machine Learning.

## 1. EXTENDED ABSTRACT

The essay correction has been a big challenge since education became available for most of the people. For instance, the Brazilian government performs since 1998 a National Exam in order to assess high school students. This exam is also known as ENEM (Exame Nacional do Ensino Médio) and evaluates different kinds of skills. The writing skill is also evaluated in ENEM, however the assessment of an essay is expensive because experts are hired to assess the essays. In its last edition, the Brazilian government hired 40% more evaluators than in previous edition [1]. Also, when the volume of essays is large it is almost infeasible to assess essays in a reliable manner.

The reliability and the volume of essays concerned educators around the world; therefore much Automatic Essay System (AES) was developed [2][3][4][5]. However, the research on AES evolved only for English essays. Also, the uniqueness of Portuguese language and the nature of ENEM should be considered when an AES is built for ENEM essays.

One of the earliest AES was developed by Larkey [2]. The Larkey's technique has two main steps. The first step does four classifications, which are performed by four binary Bayesian classifiers. As grade's essay can be between one and four, each Bayesian classifier determines whether the grades belong to the classifier class or not. For instance, the classifier one determines whether the essay is assessed as grade one or not assessed as grade one. So as to classify an essay, the classifier uses the set of terms of the essay. The

set of terms are preprocessed in two phases: the first phase removes stop *words*, and the second phase a stemmer is applied to terms. The Bayes Theorem is used to estimate the probability of an essay belongs to a class (grade). The second step performs linear regression in the results of the four binary classifiers. Hence, resulting grade is the real number approximated by Linear Regression. Larkey's technique was tested in a dataset composed by the discursive questions, which can be divided in the following categories:

- *Soc*: social studies question where certain facts were expected to be covered;
- *Phys*: physics question requiring an enumeration and discussion of different kinds of energy transformations in a particular situation);
- *Law*: required the evaluation of a legal argument presented in the question;

Results achieved an accuracy of 54% to 62% in the *Soc* set; 44% to 55% in the *Phys* set; and of 24% to 42% in the *Law* set.

In more recent research, Mayfield and Rosé [5] published a software called LightSide that performs Automatic Essays Correction. LightSide can use different machine learning techniques in order to correct English Essays, namely Naïve Bayes, Sequential Minimal Optimization (SMO), and J48. Although LightSide is a public software, Latifi et. al. [6] observed high quality results in correction of English Essays.

In regard to Portuguese Essays, IntelliMetric [7] is the only software found in literature that assesses Portuguese Essays. However, because IntelliMetric is private software it is difficult to evaluate its performance. Another researches approach Portuguese essays from a different perspective. For instance, Pardo and Nunes evaluate the discourse of Portuguese Essays and Souza and Feltrim analyze automatically the semantic coherence in Academic Abstracts written in Portuguese.

The few research found about AES for Portuguese Essays motivate the development of a new AES designed for Portuguese Essays. The first problem to develop this research was the lack of public dataset. So as to solve this issue, essays were collected from UOL Banco de Redações<sup>1</sup>. After that, the strategy to assign grades to essays was developed according to the dataset collected.

The essays grades from UOL vary from 0 to 10, considering 0.5 steps between the grades; therefore there are twenty-one possible grades for an essay. Hence, the problem of assign grades to an essay is model as a classification problem, and an essay can be labeled with one class belonging to the twenty one classes determined.

<sup>1</sup> <http://educacao.uol.com.br/bancoderedacoes/>

In order to build a bayes classifier for the Essays, the Bayes' Theorem was used. The Bayes' Theorem can be defined like the following formula.

$$P(A|B) = P(B|A) \times P(A)/P(B)$$

In the developed algorithm the Bayes' theorem can be rewritten as the following formula.

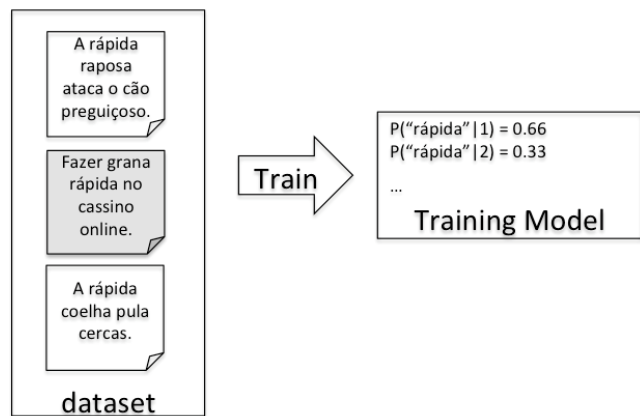
$$P(\text{grade}|\text{essay}) = P(\text{essay}|\text{grade}) \times P(\text{grade})/P(\text{essay})$$

Essay is represented like a set of terms or words, therefore the conditional probability  $P(\text{essay}|\text{grade})$  is computed using the following equation.

$$P(\text{essay}|\text{grade}) = \prod P(\text{term}|\text{grade})$$

The equation above can be explained by the idea that the probability that an essay belongs to a grade is products between the probabilities of terms from given essay belong to given category. After this computation, the probability  $P(\text{grade})$  is calculated considering a randomly selected document belongs to a given category, i.e., number of documents in given category divided by total number of documents. Since  $P(\text{document})$  is the same no matter what category the calculation is being done for, it will scale the results by the exact same amount, so you can safely ignore that term.

Training phase computed the probabilities of all terms in training set belong to a class. Figure 1 describes how our training model is built. Dataset in example has only two classes; the first essay is graded with one is a blank document, the second essay is graded with two is a gray document, and the third essay is graded with one is blank document. Training model is recorded in order to use probabilities for classify test essays.



**Figure 1: Training Model Production**

After training, testing phase is performed. In this phase, the probability  $P(\text{grade}|\text{essay})$  is computed for each grade. For instance, consider a document containing the text: “coelha rápida”. If the training model used to compute the probability  $P(1| \text{“coelha rápida”})$  is the same described in Figure 1, then the following computation is calculated:

$$P(1| \text{“coelha rápida”}) = P(1) \times P(\text{“coelha”}|1) \times P(\text{“rápida”}|1)$$

Equation above can be read like: probability of essay “coelha rápida” being graded as 1 is equal to the probability of some essay of dataset being graded as 1 times the probability of “coelha” belongs to an essay grade 1 times the probability of “rápida” belongs to a essay grade 1. In this example,  $P(1)$  is 0.66,  $P(\text{“coelha”}|1)$  is 0.5 and  $P(\text{“rápida”}|1) = 1.0$ . Therefore the value of  $P(1|\text{“coelha rápida”})$  is 0.33. Same procedure is executed when  $P(2|\text{“coelha rápida”})$  is computed, and the result of this computation is 0.

The dataset used in our experiments was collected in a web site called UOL Banco de Redações. The dataset collected is also available for download in raw text<sup>2</sup>. Essays from UOL are categorized by topics, which were proposed by experts. Each topic owns between thirteen to twenty essays. Table 1 describes quantities in corpus built.

**Table 1: Dataset Quantities**

dataset	#essays	#words
Train	379	11617
Test	50	3354

Grades assigned to essays observe the following guidelines from ENEM.

- Observe formal writing rules of Portuguese language.
- Comprehension of proposed topic and application from various human knowledge areas to develop the essay, within the structural boundaries of an argumentative essay.
- Selecting, relating, organizing and understanding information, facts, opinion and arguments in order to defend a point of view.
- Exhibiting the language knowledge that is necessary for argument construction.
- Building a proposal for solve the given problem, demonstrating respect in regard to human being values and considering social and cultural diversity.

Each of the guidelines above is graded according to range from 0 to 2. Thus, the resulting grade to essay is to sum all the subgrades assigned to essay.

Grades assigned by human evaluators is usually different, according to Page [10], Pearson Correlation between human evaluators is about 0.564. Therefore it is reasonable

to evaluate not exactly the grade automatically assigned but also adjacent grades.

**Table 2: Correlation Between Predict Grades**

Strategy	Pearson Correlation
Proposed Approach	0.396
Larkey Bayesian	0.63
Human Evaluators	0.564

Although the correlation of our proposed approach is almost 0.17 far from human evaluators, our strategy is simple and we considered it as baseline system. Also, dataset used in strategies are different.

Table 3 describes the accuracy achieved by our classifier. Adjacent grades are grades classified at most 1.5 points away from the human evaluator grade.

**Table 3: Accuracy Baseline Strategy**

Accuracy: Adjacent Grades (1.5)
52%

Unlike Larkey, our strategy does not use binary classifiers, but only one classifier. Also, grades classified by Larkey vary from 0 to 4, while grades classified in our baseline system vary from 0 to 10. These issues are reasons for lower results achieved.

However, we achieved the goal to build a first Portuguese corpus to AES and a baseline system for future comparison.

Next steps for this research comprises following items:

- Remove stop words from essay.
- Apply stemmer in words of essay.
- Build Binary Classifiers for the twenty-one classes.
- Apply Linear Regression to results of binaries classifiers.
- Use a more robust Machine Learning approach than Bayesian Classifier.

Although, there are still some future researches to do, we consider that a public Portuguese corpus and a baseline for Automatic Essay System is a contribution for future developments in Brazilian AES research.

## 2. REFERENCES

- [1] Moura, Rafael Moraes. Número de Corretores de redação do Enem cresce 40%. May 24th 2012. Available at: <http://www.estadao.com.br/noticias/vidae.numero-de-corretores-de-redacao-do-enem-cresce-40,877480,0.htm> .

<sup>2</sup>

<https://drive.google.com/folderview?id=0B35NbJbdG5JqQXcxQV9UcTdjS0k&usp=sharing>

- [2] Larkey, Leah S. "Automatic Essay Grading Using Text Categorization Techniques". In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.
- [3] Rudner, Lawrence M., and Tahung Liang. "Automated essay scoring using Bayes' theorem." *The Journal of Technology, Learning and Assessment* 1.2 (2002).
- [4] Bin, L., Jun, L., Jian-Min, Y., & Qiao-Ming, Z. "Automated Essay Scoring Using the KNN Algorithm". 2008 International Conference on Computer Science and Software Engineering, 735–738. doi:10.1109/CSSE.2008.623
- [5] Mayfield, Elijah and Rosé, Carolyn Penstein. *LightSide: Text Mining and Machine Learning User's Manual*. 2012. Carnegie Mellon University.
- [6] Latifi, Syed M. Fahad, Guo, Qi, Gierl, Mark J., Mousavi, Amin, Fung, Karen. *Towards Automated Scoring using Open-Source Technologies*. Annual Meeting of Canadian Society for the Study of Education. 2013.
- [7] Dikli, Semire. "An overview of automated scoring of essays." *The Journal of Technology, Learning and Assessment* 5.1 (2006).
- [8] Pardo, T. A. S., and Nunes, Maria das Graças Volpe Nunes. "On the development and evaluation of a Brazilian Portuguese discourse parser." *Revista de Informática Teórica e Aplicada* 15.2 (2008): 43-64.
- [9] Souza, V. M. A., and Feltrim, V. D. F.. "Automatic Analysis of Semantic Coherence in Academic Abstracts Written in Portuguese." *Methodology* 273 (2011): 11-90.
- [10] Page, Ellis Batten. "Computer grading of student prose, using modern concepts and software." *The Journal of experimental education* 62.2 (1994): 127-142.