Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task

Jeziel C. Marinho[1], Rafael T. Anchiêta[2], Raimundo S. Moura[1]

[1] Federal University of Piauí (UFPI) – Teresina, PI – Brazil
{jezielcm, rsm}@ufpi.edu.br
[2] Federal Institute of Piauí (IFPI) – Picos, PI – Brazil
rta@ifpi.edu.br

Abstract. Automatic Essay Scoring (AES) is the computer technology that evaluates and scores the written essays, aiming to provide computational models to grade essays automatically or with minimal human involvement. While there are several AES studies in a variety of languages, few of them are focused on the Portuguese language. The main reason is the lack of a corpus with manually graded essays. In order to bridge this gap, in this paper we extended a corpus of essays written by Brazilian high school students in an online platform. All of the essays are argumentative and were scored across five competences by experts. Moreover, we conducted an experiment with the extended corpus to show some challenges posed by the Portuguese language. The corpus are publicly available at `https://github.com/lplnufpi/essay-br`.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Language resources; I.7 [**Document and Text Processing**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous

Keywords: Automated Essay Evaluation, Textual Databases, Natural Language Processing

1. INTRODUCTION

The Automated Essay Scoring (AES) area began with [Page 1966] in the Project Essay Grader system, which according to [Ke and Ng 2019] remains since then. [Shermis and Barrera 2002] define AES as the computer technology that evaluates and scores the written prose, i.e., it aims to provide computational models for automatically grading essays or with minimal involvement of humans [Page 1966].

AES is one of the most important educational applications of Natural Language Processing (NLP) [Ke and Ng 2019; Beigman Klebanov et al. 2016]. It encompasses some other fields, such as Cognitive Psychology, Education Measurement, Linguistics, and Written Research [Shermis and Burstein 2013]. They aim to study methods to assist teachers in automatic assessments, providing a cheaper, faster, and deterministic approach than humans do when scoring an essay. Due to all benefits, AES has been widely studied in various languages, for example, English, Chinese, Danish, Japanese, Norwegian, and Swedish, among others [Beigman Klebanov and Madnani 2020].

To grade an essay, these studies supported the development of regression-based methods, such as [Beigman Klebanov et al. 2016; Vajjala 2018], classification-based methods as [Farra et al. 2015; Nguyen and Litman 2018], and neural networks-based methods as [Taghipour and Ng 2016]. Moreover, AES systems have also been successfully used in schools and large-scale exams [Williamson 2009]. According to [Dikli 2006], examples of such systems are: Intelligent Essay$^{\text{TM}}$, Criterion$^{\text{SM}}$, IntelliMetric$^{\text{TM}}$, E-rater®, and MY Access!®.

Despite the importance of the AES area, most of the resources and methods are only available for the English language [Ke and Ng 2019]. There are very few AES-based studies for the Brazilian Portuguese language, such as [Bazelato and Amorim 2013; Amorim and Veloso 2017; Fonseca et al. 2018]. The main reason for that is the lack of a public corpus with manually graded essays. Hence, it is important to put some effort into creating resources that will be useful for the development of alternative methods for this field.

In this paper, aiming to fulfill this gap, we extended the Essay-BR corpus [Marinho et al. 2021] with essays made available by other Brazilian researchers. These essays are of the argumentative type and were graded by experts across five different competences to reach the total score of an essay. The competences follow the evaluation criteria of the ENEM exam - ***E**xame **N**acional do **E**nsino **M**édio* - (National High School Exam), which is the main Brazilian high school exam that serves as an admission test for most universities in Brazil. The extended corpus has $2,009$ more essays than the previous corpus. Moreover, we performed a detailed analysis of the extended corpus, providing insights for the AES task.

In addition to the corpus, we carry out an experiment, implementing two approaches to automatically score essays, demonstrating the challenges posed by the corpus, and providing baseline results. It is important to highlight that the corpus of essays meets the new ENEM evaluation criteria and, we believe it will foster AES studies for the Portuguese language, resulting in the development of alternative methods to grade an essay.

This article is an extended and revised version of a previous conference paper [Marinho et al. 2021], presented in the 36th edition of the Brazilian Symposium on Databases (SBBD 2021) - Dataset Showcase Workshop (DSW). The contributions of this extended version are: i) extension of the corpus with $1,160$ new essays from Vestibular UOL and 849 essays from the dataset made available by [Amorim and Veloso 2017]; ii) classification of essays into five levels, namely: **precarious**, essays with grades from 40 to 240; insufficient, essays with grades from 200 to 440; **medium**, essays with grades from 400 to 640; **good**, essays with grades from 600 to 840; and **excellent**, the essays with grades from 800 to $1,000$; iii) inclusion of a new table with statistics regarding the essay categories, in addition to updating data from existing tables; and iv) reclassification of essay topics (prompt[1]) and inclusion of Table X with statistics, considering the new topics.

The remaining of this paper is organized as follows. Section 2 describes the main related works. In Section 3, we present the ENEM exam. Section 4 details our corpus, its construction, and an analysis of the training, development, and testing sets. In Section 5, we describe the conducted experiments. Finally, Section 6 concludes the paper, indicating future work.

2. RELATED WORK

Currently, there are few AES-based studies for the Brazilian Portuguese language. Here, we briefly present them.

[Bazelato and Amorim 2013] crawled 429 graded essays from the web site called *UOL Banco de Redações* [2] to create the first corpus of essays for the Portuguese language. However, the crawled essays are too old and do not meet the ENEM exam criteria. For example, the essay grades vary from 0 to 10, considering 0.5 steps between the grades. It is important to mention that this resource is a preliminary version of the dataset made available by [Amorim and Veloso 2017].

[Amorim and Veloso 2017] developed an automatic essay scoring method for the Brazilian Portuguese language. For that, they collected $1,840$ graded essays about 96 topics from the *UOL Essay*

[1] We used the term “prompt” to indicate the proposed theme or topic for an essay. It will be adopted throughout the text.

[2] `https://drive.google.com/folderview?id=0B35NbJbdG5JqQXcxQV9UcTdjS0k&usp=sharing`

Database website [3]. Next, they developed 19 features to feed a linear regression to grade the essays. Then, to evaluate the approach, the authors compared the automatic scores with the scores of the essays, using the Quadratic Weighted Kappa (QWK) metric [Cohen 1968], achieving 42.45%. Just as in the [Bazelato and Amorim 2013] work, the collected essays are very old and do not meet the current ENEM exam criteria, as each competence is scored according to the scale from 0 to 2 (step: 0.5), and the final score is the sum of all competence scores. In a posterior work, [Amorim et al. 2018] analyzed the presence of biased ratings in the AES area. They showed that removing biased scores from the training set results in improved AES models.

[Fonseca et al. 2018] addressed the task of automatic essay scoring in two ways. In the first one, they adopted a deep neural network architecture similar to the [Dong et al. 2017] with two Bidirectional Long Short-Term Memory (BiLSTM) layers. The first layer reads word vectors and generates sentence vectors, which are read by the second layer to produce a single essay vector. This essay vector goes through an output layer with five units and a sigmoid activation function to get an essay score. In the second approach, the authors hand-crafted 681 features to feed a regressor to grade an essay. The authors evaluated the approaches using a corpus with 56,644 graded essays and reached the best result with the second method, achieving 75.20% in the QWK metric. Although this work had used essays written in Brazilian Portuguese to evaluate their methods, the authors did not make corpus publicly available, making the development of alternative methods difficult. Moreover, each work used a different corpus, making it difficult to compare them fairly.

In English, according to [Ke and Ng 2019], there are five popular available Corpora: ICLE [Sylviane Granger and Paquot 2009], CLC-FCE [Yannakoudakis et al. 2011], Automated Student Assessment Prize (ASAP), TOEFL 11 [Blanchard et al. 2013], and AAE [Stab and Gurevych 2014]. The ASAP corpus, one of the most famous and established corpus, was released as part of a Kaggle competition in 2012, becoming widely used for holistic scoring. Furthermore, the corpus is composed by 17,450 argumentative essays and 8 prompts written by United States students from grades 7 to 10.

In what follows, we introduce the ENEM exam.

3. ENEM EXAM

The ENEM - ***E**xame **N**acional do **E**nsino **M**édio* - (National High School Exam) is actually an exam to assess the quality of high school education, which has been later re-purposed to serve also as an admission test. More than that, it is the second-largest admission test in the world after the National Higher Education Entrance Examination, the entrance examination of higher education in China. In the ENEM exam, the reviewers take into account five competences to evaluate an essay, which are:

(1) Adherence to the formal written norm of Portuguese.
(2) Conforming to the argumentative text genre and the proposed topic (prompt), to develop a text, using knowledge from different areas.
(3) Selecting, relating, organizing, and interpreting data and arguments in defense of a point of view.
(4) Using argumentative linguistic structures.
(5) Elaborating a proposal to solve the problem in question.

where each competence is graded with scores ranging from 0 to 200 in intervals of 40. These scores are organized by proficiency levels, as shown in Table I. In this table, the 200 score indicates an excellent proficiency in the field of competence, whereas the score of 0 shows ignorance in the field of competence.

[3] `https://github.com/evelinamorim/aes-pt`

Table I: Proficiency levels of the ENEM exam.

Score	Description
200	excellent proficiency
160	good mastery
120	medium dominance
80	insufficient mastery
40	precarious dominance
0	ignorance

In this way, the total score of an essay is the sum of the competence scores and may range from 0 to $1,000$. At least two reviewers grade an essay in the ENEM exam, with the final grade of each competence being the arithmetic mean between the two reviewers. If the disagreement between the reviewers' scores is greater than 80, a new reviewer is invited to grade the essay. Thus, the final grade for each competence will be the arithmetic mean between the three reviewers.

4. EXTENDED ESSAY-BR CORPUS

The Essay-BR corpus has $4,570$ argumentative essays and 86 topics, while the extended corpus contains $6,579$ argumentative documents and 151 topics (prompts). They were collected from December 2015 to August 2021, using the same Web Scraper as the previous corpus. The topics include: human rights, political issues, healthcare, cultural activities, fake news, popular movements, covid-19, and others. Also, they are annotated with scores in the five competences of the ENEM exam. Table II summarizes the Essay-BR corpus.

Table II: Summary of the extended Essay-BR corpus.

Details	Corpus
Text type	Argumentative
Writer's language level	BR students (high school)
Scoring	Holistic
Number of essays	$6,579$
Number of prompts	151
Number of competences	5
proficiency range	$[0; 200]$
proficiency scores	$0, 40, 80, 120, 160, and\ 200$
Score range	$[0; 1,000]$

The $1,840$ essays from the dataset of [Amorim and Veloso 2017] were annotated in five grades, ranging from 0 to 2 with a scale of 0.5, i.e., their scoring scheme is different from the ENEM exam score. To transform the scores from this dataset to the ENEM exam, we mapped them to the grades recommended by the ENEM assessment manual, in the following way: 0 to 0, 0.5 to 80, 1 to 120, 1.5 to 160, and 2 to 200. Besides, of these $1,840$ essays, we included only 849 essays in our corpus because they are organized into paragraphs, while the other essays are not, that is, 991 essays do not have paragraphs, escaping the structure of an argumentative-essay text.

4.1 Construction of the corpus

To create the Essay-BR corpus, we developed a Web Scraper to extract essays from two public Websites: *Vestibular UOL (Brasil Escola)* and *Educação UOL*. The Brasil Escola website is a product currently in use, maintained by the company Rede Omnia and hosted on the UOL portal[4]. The website offers a free correction service for 120 essays per month, with a different theme being proposed each

[4] https://vestibular.brasilescola.uol.com.br/banco-de-redacoes

month. *Educação UOL* is a product maintained by the UOL portal[5], but apparently it was suspended in March 2020. It offered a free correction service for only 20 essays per month, with a different theme being proposed each month.

The essays from these Websites are public, may be used for research purposes, were written by high school students, and are graded by experts following the ENEM exam criteria. However, grades 20, 50, 100, and 150 were attributed to the competences of some essays, which is not recommended by the ENEM training manual. In total, we collected 798 essays and 43 prompts from *Educação UOL*, and $4,932$ essays and 53 prompts (43 of the Essay-BR and 10 of the extended version) from *Vestibular UOL*.

After collecting the essays, we applied a preprocessing to remove HTML tags and comments from the reviews. So, the essays contain only the content written by the students. Then, we normalized the scores of the essays. Although these Websites adopt the same ENEM exam competences to evaluate the essays, they have a slightly different scoring strategy. Thus, we made some adjustments to the grades. For example, we mapped the grade 20 from websites to 40 in our corpus, the grade 50 to 80, the grade 100 to 120, and the grade 150 to 160.

Although the corpus has a holistic scoring, it also has proficiency scores. Holistic scoring technologies are commercially valuable, since they allow automatically scoring million of essays deterministically, summarizing the quality of an essay with a single score. However, it is not adequate in classroom settings, where providing students with feedback on how to improve their essays is of utmost importance [Ke and Ng 2019]. To mitigate this weakness, the Essay-BR corpus contains five competences. Thus, a competence score shows how a student should improve their essay. For example, a student who scored 40 in the first competence, i.e., adherence to the formal written norm, got feedback that it is necessary to improve their grammar.

We also present an example of the structure of our corpus, as shown in Table III. From this table, the score is the sum of the competences (C1 to C5), and the essay content is composed as a list of paragraphs. It is important to say that some essays have no title, since, in the ENEM exam, the title is not mandatory.

Table III: Example of Essay-BR corpus.

Attribute	**Value**
Prompt	covid-19
Score	720
Title	Fighting coronavirus through science
Essay content	*list of paragraphs*
C1	160
C2	160
C3	120
C4	160
C5	120

Besides the structure, we computed some statistics, using the Natural Language Toolkit [Bird 2006] and linguistics features, using Coh-Metrix-Port [Scarton et al. 2010], about the essays of the extended corpus, as depicted in Tables IV and V, respectively. In Table IV, we can see that, on average, an essay of the corpus has 4 paragraphs, and each paragraph has 2 sentences. Furthermore, the sentences are somewhat long, with an average of 30 tokens. In Table V, one can see that most essays are in the passive voice. This is because, in Portuguese, the essays should be impersonal. Also, we calculated the Flesch score that measures the readability of an essay. From that score value, the essays are compatible with the college school level. Finally, we computed some richness vocabulary metrics,

[5] https://educacao.uol.com.br/bancoderedacoes

such as hapax legomenon, which is a word that occurs only once, lexical diversity, also known as the type-token ratio, and lexical density, which is the number of lexical tokens divided by the number of all tokens.

Table IV: Statistics of the essays.

Statistic	**Essay-BR**	
	Mean	**Std**
Paragraph per essay	4.05	1.06
Sentence per essay	10.65	4.34
Sentence per paragraph	2.62	1.42
Token per essay	324.18	95.42
Token per paragraph	79.85	34.92
Token per sentence	30.42	17.40

Table V: Linguistics features.

Feature	**Corpus**
Passive voice	75%
Active voice	25%
lexical diversity	26%
lexical density	22%
Flesch score	45.06
Hapax legomenon	35.81

To facilitate the use of the corpus in NLP tasks, the extended version was organized in the same way as the Essay-BR corpus, i.e.,with proportions of 70%, 15%, and 15%, for training, development, and testing, respectively, which corresponds to 4,605, 987, and 987 essays. Aiming to choose essays with a fair distribution of scores for each split, we computed the distribution of the total score of the essays for the corpus, as depicted in Figure 1.

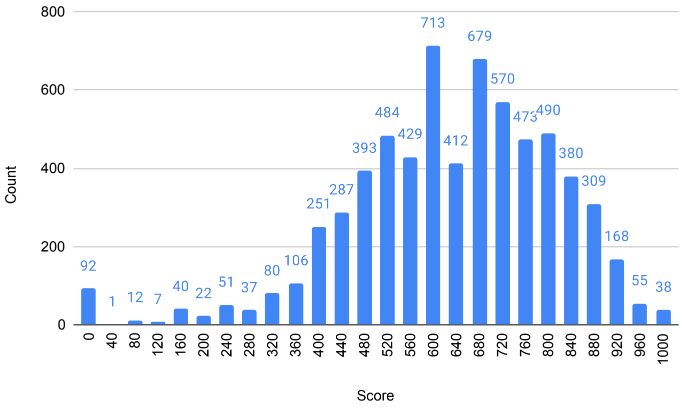

Fig. 1: Distribution of the total score.

The top 3 scores are 600, 680, and 720 corresponding to 10.83%, 10.32%, and 8.66% of the extended corpus, respectively, indicating that essays with these scores should appear more times in the training, development, and testing sets. Moreover, the scores in the corpus have a slightly rightward skewed normal distribution.

We also computed the distribution score for each competence and presented it in Table VI. Note that most essays have a grade equal to 120 in the five competences, showing that, in general, students have moderate domain in all areas.

Table VI: Distribution score for each competence.

Competence	**Scores**					
	0	40	80	**120**	160	200
C1	122	24	519	**3,136**	2,484	294
C2	123	93	918	**2,447**	2,426	572
C3	185	164	1,607	**3,055**	1,377	191
C4	207	65	886	**2,460**	1,822	1,139
C5	512	297	1,335	**2,289**	1,535	611

In the following subsection, we analyzed the training, development, and testing sets of the extended corpus.

4.2 Analysis of the extended corpus

To create the three splits with score distributions similar to that of the complete corpus, we first shuffled all the data; then, we filled each split with essays based on the score distribution. Figure 2 presents the score distribution for the training, development, and testing sets, respectively.

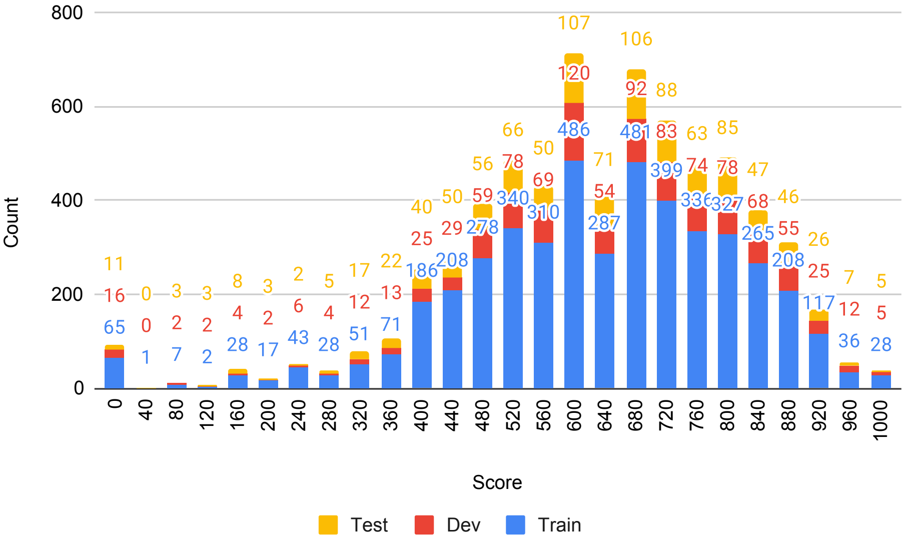

Fig. 2: Training, development, and testing sets of the extended corpus.

From this figure, one can see that the score distributions are similar to the score distribution of the complete corpus. Likewise, in the score distribution of Figure 1, the top 3 scores of the training set are 600, 680, and 720. Moreover, the development and testing sets have a similar distribution.

More than the scores, we also calculated some statistics on the splits, intending to verify whether the proportion of paragraphs, sentences, and tokens for each division remained related to the complete corpus proportion. Comparing the obtained results in Table VI with the got results of each split in Table VII, we can see that the results maintained similar proportions. For example, the average of paragraphs per essay, sentences per essay, and sentences per paragraph had related results: 4, 10, and 2, respectively.

In the following subsections, we present a discussion of the categorization of the extended corpus into five levels and the process of reclassification of the prompts (topics).

4.3 Corpus categorization

Although the ENEM exam categorize the essays by proficiency levels, as shown in Table I, we perform a different classification of the essays from the ENEM exam, as presented in Table VIII. Our intention is to provide alternatives for using the corpus in NLP tasks that deal with automatic evaluation of argumentative texts, for example, automatic correction of discursive questions in virtual learning environment (VLE).

Table VII: Statistics for each split of the extended corpus.

Split	Statistic	Mean	Standard deviation
Train	Paragraph per essay	4.06	1.09
	Sentence per essay	10.63	4.33
	Sentence per paragraph	2.61	1.42
	Token per essay	323.55	95.41
	Token per paragraph	79.60	35.13
	Token per sentence	30.42	17.34
Dev	Paragraph per essay	4.06	0.81
	Sentence per essay	10.89	4.53
	Sentence per paragraph	2.61	1.45
	Token per essay	330.39	95.42
	Token per paragraph	81.34	34.39
	Token per sentence	30.33	17.82
Test	Paragraph per essay	4.03	1.17
	Sentence per essay	10.51	4.20
	Sentence per paragraph	2.60	1.37
	Token per essay	320.88	95.25
	Token per paragraph	79.53	34.47
	Token per sentence	30.51	17.27

Table VIII: A new categorization of essays.

Category	Interval	Number #
precarious	[40; 240]	133
insufficient	[200; 440]	834
medium	[400; 640]	2,969
good	[600; 840]	3,717
excellent	[800; 1,000]	1,440

So, we create five categories: **precarious**, **insufficient**, **medium**, **good**, and **excellent**. For example, in the precarious group, the essays are in the closed interval between 40 and 240, in the insufficient group, the essays are in the closed interval between 200 and 440, and so on. It is important to note that there is an intersection among the groups. Thus, an essay scored 600, for instance, is categorized as medium and good. This new categorization may help improve results of regression (or classification)-based methods since we grouped essays in greater intervals than the original intervals from the ENEM exam. More than that, that categorization allows analyzing the differences among essays based on these groups. For example, observing the first competence (see Table IX), we may see that precarious and insufficient essay scores are lower than good and excellent essay scores. This analysis may indicate that essays with low grades in the competence C1 will have a low final grade, i.e., the candidate who fails to adherence to the formal written norm of Portuguese, will probably write a bad essay overall.

Table IX: Competences of precarious, insufficient, good, and excellent essays.

Category	Competence	Scores					
		0	40	80	120	160	200
precarious	C1	23	17	78	15	0	0
insufficient		6	19	353	417	35	4
good		1	0	18	1,627	1,944	127
excellent		0	0	0	129	1,079	232

4.4 Reclassification of the prompts

The essays of the extended corpus were analyzed by two specialists who reclassified them taking into account the topic (prompt) addressed by each one. As shown in Table X, the experts distributed the prompts into 8 groups (subcategories): Science and technology, Economics, Education, Environment, Politic, Health, Safety, Society and culture. For example, prompts such as "The culture of cancellation on social media" and "Democratization of access to cinema in Brazil" were classified in the Society and culture subcategory and prompts as "Functional illiteracy in Brazil: why this problem still persists" and "Challenges in Distance Education in Brazil" were included in the Education subcategory.

Table X: Reclassification of topics (prompts).

Subcategory	**#Prompts**	**#Essays**	**Average score**	**Standard deviation**
Science and technology	7	224	**668.92**	193.95
Economy	6	193	568.70	149.85
Education	16	606	589.70	176.86
Environment	12	427	584.07	153.70
Politics	15	338	592.42	237.27
Health	19	$1,297$	**673.30**	167.26
Safety	15	583	636.08	177.89
Society and culture	**61**	**2,909**	635.13	177.24

It is also possible to observe in Table X that the subcategory with the highest number of prompts is Society and culture (61) and, consequently, it is also the subcategory with the highest number of essays $(2,909)$. On the other hand, the subcategories with the highest average score are Health (673.30) and Science and technology (668.92). Finally, this reclassification of the prompts in the extended corpus aims to support new studies related to the topics covered in the essays.

In what follows, we present the experiment and obtained results.

5. EXPERIMENTS AND RESULTS

We carried out an experiment on the extended corpus to understand the challenges introduced by the corpus. For that, we implemented the feature-based methods of [Amorim and Veloso 2017] and [Fonseca et al. 2018]. We are aware that, in recent years, the NLP area has been dominated by the transformer architectures, as BERT [Devlin et al. 2019]. However, for the AES field the obtained results by these architectures are similar to traditional models, such as N-grams at high computation cost [Mayfield and Black 2020]. Thus, as a baseline, we preferred to implement feature-based methods since they require less computational resources and effort.

[Amorim and Veloso 2017] developed 19 features: number of grammatical errors, number of verbs, number of pronouns, and others. These features fed a linear regression to score an essay. [Fonseca et al. 2018] created a pool of 681 features, as the number of discursive markers, number of oralities, number of correct words, among others, and these features fed the gradient boosting regressor to score an essay. To extract features, we used the same tools reported by the authors, and to implement the regressors, we used the scikit-learn library [Pedregosa et al. 2011].

We evaluated those methods using the Quadratic Weighted Kappa (QWK), which is a metric commonly used to assess AES models [Yannakoudakis and Cummins 2015], and the Root Mean Squared Error (RMSE), which is a metric employed to regression problems. Table XI shows the QWK metric results, while Table XII presents the results for the RMSE metric. In the QWK metric, the greater the value, the better the result, whereas in the RMSE metric, the smaller the value, the better the result. Potential values of the QWK metric range from -1 (representing complete disagreement) to 1 (representing complete agreement). A kappa value is 0 when the expected agreement is due to chance.

Although the approach of [Fonseca et al. 2018] achieved better results in both metrics for each competence (C1 to C5), these results are not fit for summative student assessment, as usually for the AES field, threshold values between 0.6 and 0.8 QWK are used as a floor for testing purposes

[Mayfield and Black 2020]. Furthermore, the method of [Fonseca et al. 2018], which achieved 75.20% in the QWK metric in their corpus, reached only 51% in the Essay-BR. This difference may be due to two factors. The first is the size of the corpus: [Fonseca et al. 2018] used more than 50, 000 essays, whereas our corpus has 6, 579 essays. The second is implementation details: [Fonseca et al. 2018] used several lexical resources, but they did not make them available. Thus, we do not know if the lexical resources we used are the same as [Fonseca et al. 2018].

As we can see, it is necessary to develop more robust methods to grade essays for the Portuguese language in order to improve the results.

Table XI: Quadratic Weighted Kappa on the test set.

Model	C1	C2	C3	C4	C5	Total
[Amorim and Veloso 2017]	0.39	0.46	0.40	0.38	0.34	0.49
[Fonseca et al. 2018]	0.44	0.48	0.42	0.47	0.38	0.53

Table XII: Rooted Mean Squared Error on the test set.

Model	C1	C2	C3	C4	C5	Total
[Amorim and Veloso 2017]	32.26	33.45	38.20	39.35	48.18	161.09
[Fonseca et al. 2018]	32.16	33.55	38.00	38.32	47.66	157.33

In addition to these experiments, we also evaluated the essays through the prompts. We used the method of [Fonseca et al. 2018] and the subcategories Education, Health, and Society and culture, as they present more essays than the other subcategories (see Table X). We used the Quadratic Weighted Kappa (QWK) and Rooted Mean Squared Error (RMSE) metrics to measure the quality of essays. For each category, we split the essays into 90% and 10% for training and testing, respectively. Table XIII shows the results. Therefore, with the reclassification of prompts, it is possible to analyze each subcategory singly, identifying the performance of the regression models.

Table XIII: Evaluation by subcategories.

Prompt	QWK	RMSE
Education	0.43	176.31
Health	0.61	161.63
Society and culture	0.58	146.31

6. FINAL REMARKS

In this paper, we extended the Essay-BR corpus, which is a corpus written by Brazilian high school students that were graded by experts following the evaluation criteria of the ENEM exam. At this time, it has 6, 579 essays and 151 prompts but we already scraped 13, 306 essays from the *Vestibular UOL* Website. These essays are being pre-processed and will be available as soon as possible. We hope that this resource will foster the research area for Portuguese by developing of alternative methods to grade essays. More than that, according to [Ke and Ng 2019], the quality of an essay may be graded adopting different dimensions, as presented in Table XIV.

From this table, one can see that a corpus of essays may be graded regarding several dimensions. Assessing and scoring these dimensions helps the students get better feedback on their essays, supporting them to identify which aspects of the essay need improvements.

Some of these dimensions do not seem challenging, such as the grammaticality, usage and mechanism dimensions, since they already have been extensively explored. Several other dimensions, such as cohesion, coherence, thesis clarity, and persuasiveness, bring problems that involve computational

Table XIV: Dimensions to grade the quality of an essay.

Dimension	Description
Grammaticality	Grammar analysis
Usage	Use of prepositions, word usage
Mechanics	Spelling, punctuation, capitalization
Style	Word choice, sentence structure variety
Relevance	Relevance of the content to the prompt
Organization	How well the essay is structured
Development	Development of ideas with examples
Cohesion	Appropriate use of transition phrases
Coherence	Appropriate transitions between ideas
Thesis clarity	Clarity of the thesis
Persuasiveness	Convincingness of the major argument

modeling in different levels of the text. Modeling these challenging dimensions may require understanding the essay content and exploring the semantic and discourse levels of knowledge. Thus, there exist several possible applications that the Essay-BR corpus may be useful.

Finally, future works are: i) increase the corpus, which is already in process; ii) provide essay corrections, aiming to develop machine learning models to learn from the corrections; iii) make new experiments, considering the new categories of essays and groupings by reclassified prompts; and iv) analyse the influence of the year of essay writing on Automatic Essay Scoring systems.

ACKNOWLEDGMENT

The authors are grateful to IFMA for supporting this work.

REFERENCES

Amorim, E., Cançado, M., and Veloso, A. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pp. 229–237, 2018.

Amorim, E. and Veloso, A. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pp. 94–102, 2017.

Bazelato, B. and Amorim, E. A bayesian classifier to automatic correction of portuguese essays. In *XVIII Conferência Internacional sobre Informática na Educação*. Nuevas Ideas en Informática Educativa, Porto Alegre, Brazil, pp. 779–782, 2013.

Beigman Klebanov, B., Flor, M., and Gyawali, B. Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, San Diego, CA, pp. 63–72, 2016.

Beigman Klebanov, B. and Madnani, N. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pp. 7796–7810, 2020.

Bird, S. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Association for Computational Linguistics, Sydney, Australia, pp. 69–72, 2006.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. Toefl11: A corpus of non-native english. *ETS Research Report Series* 2013 (2): i–15, 2013.

Cohen, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70 (4): 213–220, 1968.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186, 2019.

Dikli, S. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment* 5 (1): 1–36, 2006.

Dong, F., Zhang, Y., and Yang, J. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, pp. 153–162, 2017.

Farra, N., Somasundaran, S., and Burstein, J. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado, pp. 64–74, 2015.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. Automatically grading brazilian student essays. In *Proceedings of the 13th International Conference on Computational Processing of the Portuguese Language*. Springer International Publishing, Canela, Brazil, pp. 170–179, 2018.

Ke, Z. and Ng, V. Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, Macao, China, pp. 6300–6308, 2019.

Marinho, J. C., Anchiêta, R. T., and Moura, R. S. Essay-br: a brazilian corpus of essays. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2021 Companion*. SBC, Online, pp. 53–64, 2021.

Mayfield, E. and Black, A. W. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Online, pp. 151–162, 2020.

Nguyen, H. V. and Litman, D. J. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, Louisiana, USA, pp. 5892–5899, 2018.

Page, E. B. The imminence of... grading essays by computer. *The Phi Delta Kappan* 47 (5): 238–243, 1966.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* vol. 12, pp. 2825–2830, 2011.

Scarton, C., Gasperin, C., and Aluisio, S. Revisiting the readability assessment of texts in portuguese. In *Proceedings of the 12th Ibero-American Conference on Artificial Intelligence*. Springer, Bahía Blanca, Argentina, pp. 306–315, 2010.

Shermis, M. D. and Barrera, F. D. Exit assessments: Evaluating writing ability through automated essay scoring, 2002.

Shermis, M. D. and Burstein, J. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

Stab, C. and Gurevych, I. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp. 1501–1510, 2014.

Sylviane Granger, Estelle Dagneaux, F. M. and Paquot, M. *International Corpus of Learner English (Version 2)*. UCL Presses de Louvain, 2009.

Taghipour, K. and Ng, H. T. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pp. 1882–1891, 2016.

Vajjala, S. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* 28 (1): 79–105, 2018.

Williamson, D. M. A framework for Implementing Automated Scoring. In *Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education*. San Diego, CA, pp. 39, 2009.

Yannakoudakis, H., Briscoe, T., and Medlock, B. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 180–189, 2011.

Yannakoudakis, H. and Cummins, R. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado, pp. 213–223, 2015.