A New Benchmark for Automatic Essay Scoring in Portuguese

Igor Cataneo Silveira and **André Barbosa** and **Denis Deratani Mauá**
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
{igorcs, aborbosa, ddm}@ime.usp.br

Abstract

Automatic Essay Scoring promises to scale up student feedback of written input, considerably improving learning. Resources for Automatic Essay Scoring in Portuguese are however scarce, not publicly available or contain inaccuracies that degrade performance. Moreover, they lack data provenance and a richer annotation and analysis. In this work we mitigate those issues by presenting a new benchmark for the task in Brazilian Portuguese. We accomplish that by downloading a collection of publicly available essays from websites that simulate University Entrance Exams, making both processed and raw data available, having a subset of the essays graded by expert annotators to assess the quality and difficulty of the task, and carrying out an extensive empirical analysis of state-of-the-art predictors considering multiple evaluation criteria.

1 Introduction

Grading essays is a ubiquitous and crucial task in Education. For the instructor, the task consumes valuable time and effort in both the grading process per se and in training and preparation (especially for junior teachers and assistants or in standardized exams). For the student, having adequate and timely feedback is essential to correct misunderstandings, encourage reflection, support engagement and maintain trust in the evaluation process.

While the importance of both scoring and commenting (i.e., providing feedback in written form) has been stressed since Page (1966)'s seminal work, most research and technological developments have focused on the scoring aspect, known as Automatic Essay Scoring (AES).

AES systems are now widespread (Beigman Klebanov and Madnani, 2021); popular standardized exams such as TOEFL, GMAT, GRE and PTE all rely on some form of AES (Attali and Burstein, 2006; Beigman Klebanov and Madnani, 2020). In addition to English, there are AES systems for a large variety of languages such as French (Lemaire and Dessus, 2003), Danish, Finnish (Beigman Klebanov and Madnani, 2020), Chinese (Song et al., 2016), Arabic (Mezher and Omar, 2016) and Japanese (Ishioka and Kameda, 2006), to name a few.

AES systems for (Brazilian) Portuguese have been developed by Amorim and Veloso (2017); Fonseca et al. (2018); Marinho et al. (2021). They are variously based on training Machine Learning models from corpora of human-annotated essays. The data sources are web sites and platforms used by high-school students for practicing for University Admission Exams, where students submit essays in exchange of feedback in the form of scores and comments. While important, those systems fall short of providing a good benchmark for AES in Portuguese, for the following reasons.

The annotated essays in the work of Amorim and Veloso (2017) were graded using a scale different from the the standardized exam it attempts to simulate, and contains no information about the scoring guidelines used by annotators. This makes it difficulty to enlarge the dataset with new essays and to validate or assess annotations. The very large data used by Fonseca et al. (2018) are proprietary and were not made publicly available. The Essay-Br corpus, used by Marinho et al. (2021), despite being relatively large and accessible, has many shortcomings. First, the HTML sources were not properly parsed to strip out unwanted content, which resulted in having annotator comments appearing in the middle of the text, ill-formed sentences, and artificial artifacts such as blank spaces and noticeable marks where comments appeared in the HTML source. That can artificially boost a machine-learning approach performance by data leakage as well as hurt the system's performance due to noisy input. Second, there was no analysis of the quality of the annotations provided, nor of

the consistency and adequacy of themes and form of essay proposals. Finally, the baseline evaluation reported was limited in terms of criteria that can be used to analyze (automatic) grading of such standardized essays, which is often a multidimensional evaluation.

This work fills the gaps in AES benchmarking for Brazilian Portuguese by:

- presenting and releasing a carefully built corpus of human-graded essays downloaded from the same sources of Essay-Br while making available also the HTML sources,
- analyzing the quality of annotations, themes and sources of the data, and
- providing a more comprehensive evaluation of state-of-the-art AES methods using standard machine learning methodology and multidimensional criteria adopted in official standardized exams.

The last item was carried out by collecting additional scoring and feedback of two experienced human annotators in a subset of the texts, which also allowed us to evaluate the difficulty of the task as measured by the inter-agreement rate between annotators. All the data and code used are available at: https://github.com/kamel-usp/aes_enem.

The rest of the paper is organized as follows. We present in Section 2 some details about the form and grading guidelines of the ENEM exam; simulating that exam is the objective of the websites from which collect data. Metrics for evaluating AES systems are discussed in Section 3. Then in Section 4 we review related work on AES for Portuguese. Details about the construction and a analysis of our corpus are presented in Section 5. The methods used to benchmark our corpus are described in Section 6 and the results of their evaluation are shown in Section 7. Final remarks and a summary of our contributions appear in Section 8.

2 ENEM Essays

The ENEM, short for *Exame Nacional do Ensino Médio*, is a entrance exam for higher education used as part of the selection process by the vast majority Brazilian universities, including the most prestigious institutions of the country. That makes websites that offer feedback on "ENEM-like" essay exams appealing to many students seeking higher education. We now review the form and grading strategies used in ENEM, as they are reflected on the data that we collected, as explained later.

The ENEM consists of a set of multiple-choice questions about a variety of topics (Hard Sciences, Languages, etc) and an argumentative essay. The latter part, which is our focus here, consists of a prompt on a selected topic, along with one or more supporting texts.

All essays are graded by at least two and at most four evaluators, depending on the inter-agreement rate. Each evaluator provides a score of 0, 40, 80, 120, 160 or 200 relative to five different competencies: fluency, writing style, argumentation quality, proper use of textual connectors, and quality of the solution to the prompt's problem. An overall score is obtained as the sum of all the competence scores. Two evaluators are considered divergent if their overall score differs by more than 100 points or if their scores differ by more than 80 points for some competence. If two evaluators are divergent, the essay is evaluated by third person, and, if still a divergence is found, by a fourth evaluator.

The evaluators of the official exam are experienced professionals and receive specialized training before grading. The training involves objective guidelines about each competence and aims at reducing disagreement. Such guidelines may vary but are generally consistent. In 2019, the Grader's Handbook, containing such guidelines, was made public for the first time.[1] Those guidelines heavily influenced this work.

3 AES Evaluation

Essay Scoring is generally posed as an ordinal regression task (McCullagh, 1980; Li and Lin, 2007), that is, the output is a finite set of ordered values such as bad < neutral < good, or, as in the official per-competence ENEM scoring rule, $0 < 40 < 80 < 120 < 160 < 200$. It is also possible to pose Essay Scoring as a type of interval regression, where the numbers actually indicate equal-sized intervals in which the true score falls (such as 0–40, 40–80, etc). We do not pursue this interpretation here, as in our experience human annotators tend to understand the scale in more categorical terms.

The Quadratic Weighted Kappa Coefficient (QWK) is a common measure of the level of agreement between two annotators that assign discrete

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

scores to the same objects (Cohen, 1968; de la Torre et al., 2018). The metric ranges from -1, representing complete disagreement, to 1, representing complete agreement. A value of 0 represents a agreement by chance. Typically, a value lower than 0.6 is understood as weak agreement, and values higher than 0.8 are understood as strong agreement (McHugh, 2012). QWK can be used to evaluate ordinal regressors (thus AES systems) by measuring the agreement to a ground-truth annotation (de la Torre et al., 2018).

Typical Machine Learning approaches are most often evaluated either by accuracy or some similar or derived metric (e.g. AUC, cross-entropy), when they deal with unordered classification, or by Rooted Mean Squared Error (RMSE) or Mean Absolute Error, when they deal with continuous values (regression). Ordinal Regression, thus Essay Scoring, can be easily cast into either approach and evaluated accordingly (McCullagh, 1980). While such metrics have the benefit of being simpler and easier to interpret than QWK, they ignore the idiosyncrasies of an ordinal regression task, and are very sensitive to class imbalance. QWK on the other hand, is sensitive to asymmetries in class distribution, and can lead to misleading conclusions in such cases (Yang et al., 2022). Hence, a more judicious and multiaspect evaluation should jointly take into account QWK and other typical machine learning metrics such Accuracy and RMSE.

4 AES Systems for Brazilian Portuguese

Using a corpus of 1840 human-annotated essays obtained from websites that simulate the ENEM essay exam, Amorim and Veloso (2017) developed a machine-learning AES system for Brazilian Portuguese based on handcrafted features. They evaluated their system w.r.t. both per-competence and overall scores and reported QWK values ranging from 0.13 to 0.31 in the per-competence scores and 0.36 in the overall score. Notably, at that time, the websites from which they collected their data scored each of the five competencies on a five-point scale from 0 to 2 with 0.5-point steps (instead of the six-point scale used by ENEM).

Following a similar approach, Fonseca et al. (2018) collected 56k essays on a private online platform that simulates the ENEM essay exam. They compared two types of machine learning AES methods: one consisting of handcrafted features (improved w.r.t. the work by Amorim and Veloso (2017)) and one based on deep neural nets using either GloVe vectors or Bi-LSTMs. The per-competence scores produced by deep neural nets obtained QWK values from 0.5 to 0.63, while overall scores had QWK of 0.74. The handcrafted feature method obtained QWK values that ranged from 0.5 to 0.67 for the per-competence scores and 0.75 for the overall score. Accordingly, they concluded that deep learning methods were not as effective, as they obtained similar scores with higher computational costs. The dataset used was not made public.

The Essay-Br Corpus (Marinho et al., 2021) is a publicly available dataset of 4572 essays and scores scrapped from ENEM essay exam simulator websites. The authors evaluated the same techniques in (Amorim and Veloso, 2017) w.r.t. QWK and RMSE. The per-competence QWK values ranged from 0.34 to 0.46 while the overall score achieved a QWK of 0.51. While predicting the overall score might seem easier, the per-competencies RMSE ranged from 34.16 to 49.09, and the overall score had RMSE values of 159 and 163. The corpus was later augmented to 6579 essays, but authors reported that the increase in data size did not improve performance significantly (Marinho et al., 2022).

As already discussed in the introduction, the Essay-Br presents many issues, among which the improper parsing of the HTML sources that resulted in leaked data from the annotator's comment and ill-formed sentences. Other issues include non-standardization of the prompts (sometimes they are presented as itemized lists, sometimes as simple strings), lack of supporting texts available on the original website (and to which the student and annotators had access), non-uniform re-scaling of scores to match ENEM scales, and lack of data provenance linking the texts to the original web pages or HTML sources. The last point is particularly important as the source websites are in constant change, and the annotators are likely to vary from time to time; that likely introduced a distribution shift in the data. Finally, the quality of data was never evaluated by experts with regards to the scoring and the similarity of themes and format to the exam they attempt to replicate (i.e., ENEM).

Sirotheau et al. (2021) compared automatic scoring and human-made scoring using a corpus of essays written in Brazilian Portuguese as part of public hiring processes. The corpus was not released publicly. Each essay was graded by at least two annotators. A random forest classifier with

140 handcraft features was learned in a supervised fashion, although the authors did not inform how the annotations were used for that purpose (since each essay has more than one, possibly disagreeing label). By using QWK, the authors concluded that the trained model had a higher inter-agreement rate with human scoring than that of human annotators.

5 A New Corpus for AES in Portuguese

In this section, we present the methodology we used to collect and annotate the new dataset, as well as relevant statistical analysis.

5.1 Data Collection

In order to mitigate the issues with the previous corpora, we developed a new dataset of essays extracted from websites that simulate the ENEM essay exam. We extracted data from the same websites used in Essay-Br, namely, *Educação UOL*[2] and *Brasil Escola*[3]. We call them Source A and Source B, respectively, in the following.

Source A had 860 essays available from August 2015 to March 2020. For each month of that period, a new prompt together with supporting texts were given and the graded essays from the previous month were made available. Of the 56 prompts, 12 had no associated essays available (at the time of download). Additionally, there were 3 prompts that asked for a text in the format of a letter. We removed those 15 prompts and associated texts from the corpus. For an unknown reason, 414 of the essays were graded using a five-point scale of either $\{0, 50, 100, 150, 200\}$ or its scaled-down version going from 0 to 2. To avoid introducing bias, we also discarded such instances, resulting in a dataset of 386 annotated essays with prompts and supporting texts (with each component being clearly identified). Some of the essays used a six-point scale with 20 points instead of 40 points as the second class. As we believe this introduces minimum bias, we kept such essays and relabeled class 20 as class 40. The original data contains comments from the annotators explaining their per-competence scores. They are included in our dataset.

Source B is very similar to Source A: a new prompt and supporting texts are made available every month together with the graded essays submitted in the previous month. We downloaded HTML sources from 7700 essays from May 2009 to May 2023. Essays released prior to June 2016 were graded on a five-point scale, and consequently discarded. That resulted in a corpus of 3200 graded essays on 83 different prompts. Although in principle Source B also provides supporting texts for students, at the time the data was downloaded, none of them were available. To mitigate that, we extracted supporting texts from the Essay-Br corpus, whenever possible, by manually matching prompts between the two corpora. We ended up with 1000 essays containing both prompt and supporting texts and 2200 essays containing only the respective prompt. Unlike Source A, Source B contains general feedback comments for each essay, which we also include in our dataset.

To sum up, we collected and released a dataset of 3,586 graded ENEM-like essays and the respective prompts, of which 1,386 contain also supporting texts. Each instance of our final dataset contains information about its source (A or B), prompt, (possibly empty) supporting texts, essay's text, per competence scores, overall score, (possibly empty) general feedback comment, and (possibly empty) per-competence feedback comments.

5.2 Analysis

An important question is how similar the data of the two sources are, in terms of the texts (prompts and essays) and of the scoring strategies. To address the latter question, we show the histograms of per-competence scores in Figure 1. We see that the distribution of scores of Source A follows a more symmetric, bell-shaped curve, while the distribution of scores of Source B is skewed towards high scores. Source A also presents more similar distributions for all competences, while for Source B the distributions are markedly different across competences. As a comparison, Figure 2 shows the score distribution of essay grades for the ENEM 2022 exam.[4] Similar shaped distributions are observed for the years of 2019–2021. One notes that the Source A grade distribution follows more closely the real exam grade distributions, suggesting that Source B has a label-bias problem. We speculate that the difference is either due to a selection bias caused by low-quality essays being not submitted or not graded in Source B, or due to a grading strategy that inflates scores in Source B.

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

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

[4] Extracted from https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

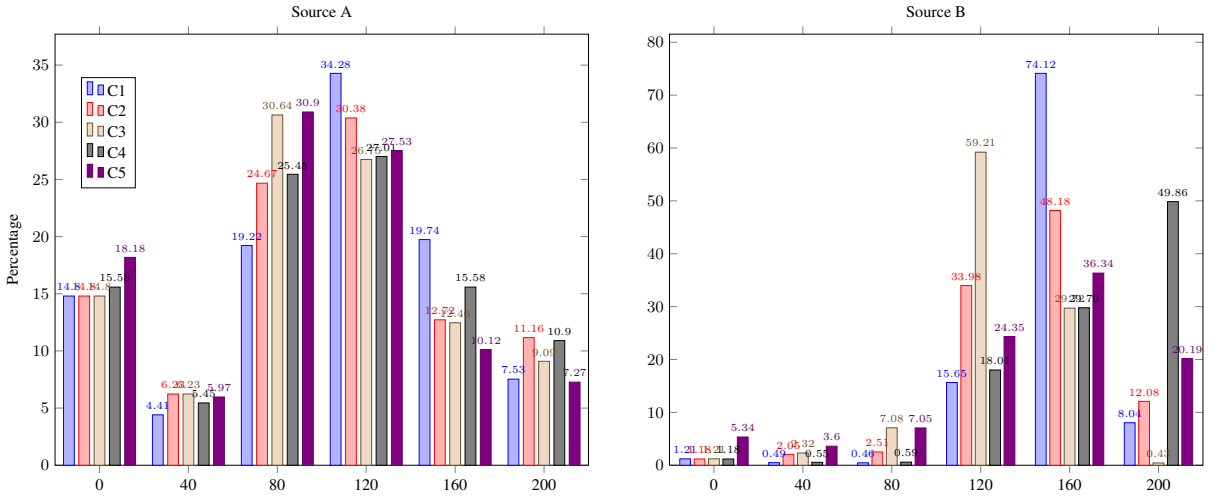

Figure 1: Per-competence score distributions of datasets.

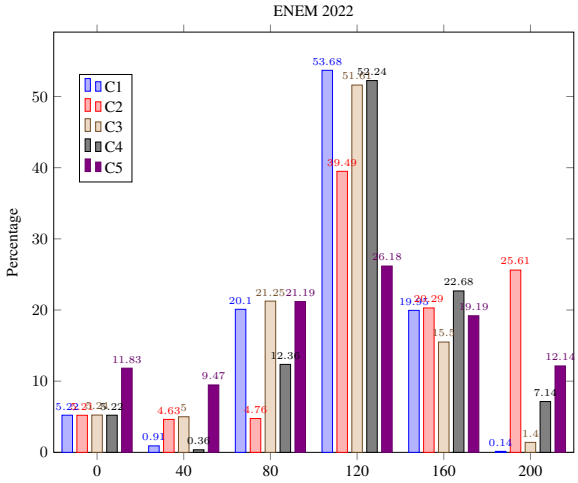

Figure 2: Per-competence score distributions in the 2022 ENEM exam.

To analyze the quality of the annotations of the data sources, we asked two experienced annotators to annotate all essays in Source A following the guidelines of ENEM 2019 Grader's Handbook. We instructed each annotator to work independently and to not communicate with the other annotator regarding the grading, since the annotators were acquainted with each other. They only discussed what would constitute tangent themes to each prompt. We denote the anonymous graders as Grader A and B, and compare their annotations against the baseline (i.e., grades from Source A) and against each other. The results appear in Table 1. Accuracy (ACC) measures the exact agreement rate. Following the ENEM guidelines, the per-competence divergence (DIV) refers to the percentage of instances where the per-competence score differed from the reference by more than 80 points. The overall divergence is the percentage of instances where the overall score differed by more than 100 points.

We notice from the tables that the inter-rater agreement between each grader and the baseline (the scores from the source data) is always in the fair-moderate range ($0.2 \leq$QWK≤ 0.6), while the inter-rater agreement between the two graders is on the moderate-substantial range ($0.4 \leq$QWK≤ 0.8), with Grader B showing a higher agreement with the baseline. Regarding the DIV column, we see that the per-competence divergence is relatively low, while the overall divergence is high; this happens because an overall divergence is not necessarily implied by a per-competence divergence.

Overall, we see that the graders show a much higher agreement between themselves than relative to the baseline. That can be explained by presuming that the baseline is actually taken from many different annotators, each partially disagreeing from each other. Note how the different metrics provide different information. RMSE and QWK are sensitive to large differences between annotations, while ACC and DIV capture total or partial agreement, respectively, and are thus less sensitive to large deviations in scores.

All things considered, we concluded that while there is significant uncertainty (or noise) in the baseline annotations of Source A, the uncertainty is consistent with human inter-rater disagreement and is informative enough to support data-based AES systems.

We incorporate the annotations of Grader A and B in our dataset. That creates an important and to

	Grader A Vs. Baseline				Grader B Vs. Baseline				Grader A Vs. Grader B			
	ACC	RMSE	QWK	DIV	ACC	RMSE	QWK	DIV	ACC	RMSE	QWK	DIV
C1	29.1	63.00	0.35	12.7	31.2	57.73	0.37	9.6	55.6	31.25	0.57	0.5
C2	23.4	71.29	0.31	15.6	26.2	54.43	0.48	7.5	45.2	48.76	0.54	4.4
C3	23.1	56.97	0.42	8.3	28.1	57.52	0.48	7.0	43.6	43.68	0.59	4.2
C4	27.0	63.88	0.27	14.5	28.1	60.85	0.37	12.5	54.5	33.06	0.45	0.8
C5	26.2	71.87	0.24	14.8	22.6	72.45	0.26	14.0	43.4	50.84	0.64	6.8
Overall	4.9	264.40	0.39	72.2	7.5	237.55	0.49	66.2	13.2	128.40	0.69	37.1

Table 1: Pairwise relative performances of Graders A, B and baseline (taken from website). ACC: % Accuracy, RMSE: Rooted Mean Squared Error, QWK: Quadratic Weighted Kappa, DIV: % of divergent instances.

our knowledge unique feature of the corpus: the ability to investigate the performance of AES systems against a set of carefully annotated essays from two different human annotators that differ from the (possibly non-curated set of) baseline annotators.

To investigate the quality of the prompts and essays, we interviewed the graders after they submitted their annotations. The graders judged that relative to the official ENEM, the topics of the essay prompts in our dataset are more controversial, more open-ended, do not explicitly ask for an intervention, and ask more than one question. This makes the essay harder for students, as it becomes necessary to connect more information. It also makes grading more challenging, as it is harder to identify tangential arguments. Regarding the written feedback available, the annotators reported finding them rude from a teacher's viewpoint.

To analyze if the text distribution is different in each source, we evaluated the performance of a domain classifier that predicts whether a given text comes from either Source A or Source B. We carried out an experiment by sampling 270 essays from each source for training and 115 for testing. To avoid data leakage, we always put essays about the same prompt in the same split. Then, we trained a neural network classifier for five epochs. We resample and rerun the experiment 20 times, which sums up to 100 tests. The domain classifier had an average accuracy of 64.57%, which lead us to conclude that the essays from different sources are similar enough. The above chance accuracy of the classifier might result from clues like the quality of the essays, given that essays from Source B have in general higher scores.

6 Baseline Methods

To establish a benchmark, we developed multiple neural network predictors based on the BERTimbau transformer model for Brazilian Portuguese (Souza et al., 2020), which comes in two variants. The base variant, with 108 million parameters, reduces overfitting risks due to our limited labeled data. The large variant, with 334 million parameters, is potentially better for capturing complex text relations. With the same architecture and variant, we obtain different predictors by using different framings of AES: a classifier (trained using cross-entropy), a regressor (trained using MSE) and a ordinal regressor (trained using CORN loss (Shi et al., 2021)).

	ACC	RMSE	QWK	Div.
Ordinal	0	3	2	2.5
Regressor	1	2	1	2
Classifier	4	0	2	0.5

Table 2: Number of times each predictor got the best performance in each metric across all competences. Points for ties are distributed by the number of predictors tied.

We compare the BERTimbau-base models against the handcrafted feature-based linear regressor described in (Amorim and Veloso, 2017) and implemented by the authors of Marinho et al. (2021). We call he latter method Handcrafted in the following. We also compare against the Zero Rule algorithm, which predicts the most frequent label in the training set.

We used each source with a different purpose. Source A was split into training, validation, and test sets, using stratification by prompt, that is, essays for the same prompt are in the same split. Additionally, we treated each annotation (baseline, Grader A and B) as a different instance. That gave us 738 instances for training, 204 for validation, and 213 for testing.

Given the discrepancy in label distribution between Sources A and B, and the lack of validated annotations for Source B, we opted to use this data

	C1	C2	C3	C4	C5
ACC	No/Yes	No/Yes	Yes/Tie	Yes/No	Yes/No
RMSE	No/Yes	No/Yes	Yes/Yes	No/Yes	Yes/Yes
QWK	No/No	Yes/No	Yes/Yes	No/Yes	Yes/Yes
DIV	No/No	Yes/No	No/Yes	No/Yes	Yes/No

Table 3: Does Method B outperforms Method MLM? Answers for base model/large model.

separately and prior to training on Source A data. We split the data from Source B randomly, using 90% for training and the 10% remaining for validation. We tested two pre-training strategies. One that disregards labels (score) and uses Masked Language Modeling (MLM) with the AdamW optimizer (Loshchilov and Hutter, 2019), monitored by a perplexity-based early stopping mechanism on the validation subset. Batch sizes were fixed at 16, using gradient accumulation if necessary, and a Learning Rate finder algorithm (Smith, 2015) determined the rates. The other strategy was supervised training through ordinal regression, targeting QWK metrics for early stopping, and learning rate fixed at 10^{-4}. We call the first strategy of Method MLM and the second of Method B. The pre-trained models were then fine-tuned on training portion of Source A, using the following hyperparameters: batches of size 16, learning rate of 10^{-4}, weight decay of 0.01, and early stopping criteria based on QWK improvements over three epochs.

7 Empirical Analysis

We first evaluate which prediction approach is best for AES: classification, regression or ordinal regression. To minimize the factors of variability, we use only data from Source A. In Table 2, we show how often a predictor type performed best in each metric across all competencies. Surprisingly, Ordinal (regression) performs best w.r.t. RMSE, despite this metric being optimized by Regressor. On the other hand, Classifier is far superior w.r.t. accuracy, where ordinal is always outperformed. Classifier is also surprisingly effective for QWK, despite disregarding the order among classes. None of the methods were optimized for divergence, and for that metric, we observe that the ordinal regressor showed superior performance. Overall, we conclude that ordinal regression outperforms the other approaches, especially when QWK and divergence are prioritized.

Next, we address whether using a larger model improves performance. For that, we trained a second version of the previous models using the large version of BERTimbau. The base and large variants tied 4 times; in 18 cases the large model was the winner, and in 38 scenarios, the base was the winner. We conclude that just increasing the size of the model does not lead to better performance for this task, possibly due to the modest data size.

To assess whether Ordinal-base is state-of-the-art, we present in Figure 3 a radar plot showing the its performance along with performances of Handcrafted and ZeroRule. The values were normalized so that 1 represents the best performance among them for each competence. For metrics where lower values are better, it was first taken their inverse. We can take that the Zero Rule algorithm is, in general, inferior to the others, but it still performs better than the others for some metric in some competence. The results vary greatly by competence and metric. Notably, we observe that ZeroRule often performs best or second-best w.r.t. ACC, which suggests a low predictive power of other methods in that regard. Ordinal performs similarly to Handcrafted in 6 cases and outperforms it in other 8 cases. Handcrafted is particularly performing for Competence C1 and C4 w.r.t. QWK and DIV; Ordinal is particularly performing for C5, and the difference between both is marginal for C2 and C3. We thus conclude that there is no clear winner, and still room for improvement.

Until now, analyses have been restricted to essays from Source A. We extend the investigation to Source B, by training an ordinal regressor using either Method MLM and Method B, as described in the previous section. The results, shown in Table 3, demonstrate that for the base-variant of BERTimbau, both approaches perform similarly, with the same number of wins each, while Method B was slightly superior for the large variant.

In light of all those results, we proceeded to a more extensive comparison the most competitive strategies: Ordinal trained on Source A with the base variant, and Method MLM with the base variant and Method B with either the base or the large variant. Those models represent the minimal, medium and maximum model complexities.

The results appear in the left part of Table 4 (Complete Test Set). We see that no strategy is consistently superior nor inferior in all competences. When we check the best performance per metric across all competences, Method B large was the best performing in 9.5 cases (one tie), followed by Method B base in 5 cases, Ordinal was the best 4.5 times and Method MLM was the best only once.

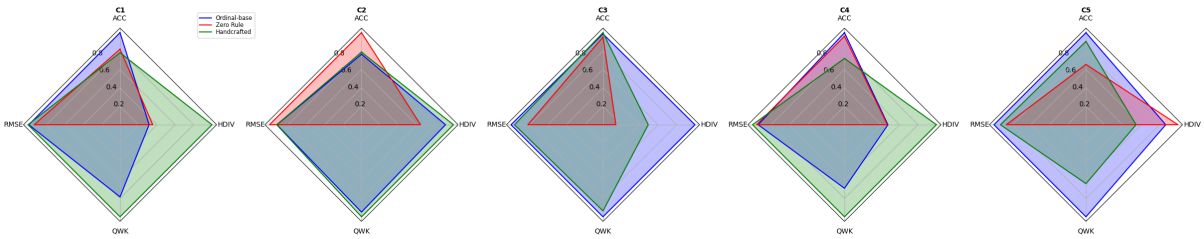

Figure 3: Comparison of BERTimbau-based ordinal regressor, feature-based linear regression and ZeroRule.

			Complete Test Set				Non-Divergent Test Set			
	Model	Size	ACC	RMSE	QWK	DIV	ACC	RMSE	QWK	DIV
C1	Ordinal	Base	48.61	44.47	0.29	7.40	54.16	31.62	0.42	1.19
	Method MLM	Base	51.85	43.46	0.33	6.94	55.95	30.39	0.49	0.59
	Method B	Base	44.90	47.76	0.29	7.40	47.02	34.64	0.43	1.19
	Method B	Large	52.31	42.68	0.37	6.48	55.95	29.11	0.55	0.00
C2	Ordinal	Base	30.50	51.35	0.37	4.62	32.73	46.39	0.44	2.38
	Method MLM	Base	34.72	53.67	0.32	7.87	36.30	48.50	0.38	4.76
	Method B	Base	27.31	56.50	0.33	5.09	30.35	52.91	0.38	3.57
	Method B	Large	38.88	54.70	0.23	8.33	41.66	50.33	0.26	5.95
C3	Ordinal	Base	29.16	46.26	0.47	0.92	31.34	45.14	0.48	0.95
	Method MLM	Base	33.33	45.21	0.42	2.77	32.83	43.61	0.45	1.99
	Method B	Base	37.96	43.96	0.46	3.70	39.30	42.41	0.47	3.48
	Method B	Large	37.03	44.88	0.50	3.70	38.30	42.97	0.52	2.98
C4	Ordinal	Base	46.29	47.45	0.29	6.94	49.09	34.35	0.33	0.00
	Method MLM	Base	38.88	42.94	0.39	2.77	42.42	33.96	0.38	0.00
	Method B	Base	53.70	45.13	0.28	8.33	55.75	30.66	0.37	0.00
	Method B	Large	45.37	41.54	0.42	3.70	49.09	30.51	0.44	0.00
C5	Ordinal	Base	30.09	51.49	0.50	3.24	32.22	47.79	0.57	1.66
	Method MLM	Base	23.61	54.36	0.26	4.16	23.33	52.66	0.29	3.33
	Method B	Base	31.94	50.18	0.53	3.70	33.88	46.85	0.59	2.22
	Method B	Large	26.85	46.98	0.50	3.24	27.22	43.10	0.58	1.66

Table 4: Performance of selected algorithms.

The good performance of Method B shows that the large model pays off when allied with more data and that Source B can be leveraged to improve performance. Although the large version has a good performance, it has a high computation cost, and, arguably, even the smaller predictor suffers from overfitting. Hence, the benefits of using a bigger model are still not completely paying off.

Finally, as some essays had divergent annotations even between humans (according to ENEM scoring guidelines), we compared the performance of predictors on the subset of essays where none of the three annotations diverged. The results are presented in the right part of Table 4 (Non-Divergent Test Set). In all competences we had a model with less than 2.5% of DIV and the highest value for this metric was 5.95% in C2. Importantly, in most cases, performances improve significantly. This shows that inter-rater disagreement and annotations collected from web sources can hurt performance and make overall evaluation difficult. We thus recommend that the Non-Divergent Test Set be used as gold standard for future evaluation.

8 Conclusion

The field of Automatic Essay Scoring (AES) can have a deep impact on education by unburdening teachers and making educational tools available to those who need them. Despite this, there are few resources for Portuguese AES. The existing research either lacks availability or a thorough evaluation.

In this work, we presented a benchmark that gathers 3586 essays from two websites (called Sources A and B) previously used in the literature

and makes them available with their HTML source. Source A essays were then scored on five different competences (traits) by two experienced annotators. We noted that agreement between either annotator and the original scores is significantly lower than inter-rater agreement, which shows that scores found online might be noisy, unreliable or inconsistent. We also analyzed the similarity of instances from either sources using a domain classifier and score distributions; we concluded that texts from the sources appear to be similar while their score distributions is markedly different.

Finally, we developed neural network predictors in order to establish a baseline for performance on the benchmark. First, we showed evidence that AES is, indeed, better framed as an ordinal regression task than classification or regression. We also experimented with different variants of BERT models for Portuguese, and concluded that larger models do not obtain superior performance, likely due to our insufficient dataset size. We also observe that BERT-based models perform slightly better than feature-based linear regressions. Finally, we showed that, despite the discrepancy between sources, using data from Source B improves performance on Source A, and that performance is maximized on the portion of data where inter-rate agreement is maximum. Our best performing models obtain per-competence quadratic weighted Kappa values between 0.26 to 0.59 for that subset. Using the same metric of standardized exam the data sources simulate, the methods achieved performances comparable to human annotators.

Our results show that there is much room for improvement. The feature-based linear regressor, while simple, was competitive for some competences; designing better features can possibly lead to state-of-the-art performance. It is also interesting to explore approaches that combine feature-based and BERT-based predictors.

Acknowledgements

This work was partially supported by FAPESP grants no. 2022/02937-9 and 2019/07665-4, CNPq grant no. 305136/2022-4 and CAPES Finance Code 001.

References

Evelin Amorim and Adriano Veloso. 2017. 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*, pages 94–102.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).

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

Beata Beigman Klebanov and Nitin Madnani. 2021. *Automated Essay Scoring*. Springer Cham.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scales disagreement of partial credit. *Psychological Bulletin*, 70:213–220.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, pages 144–154. Machine Learning and Applications in Artificial Intelligence.

Erick Rocha Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. Automatically grading brazilian student essays. In *Proceedings of International Conference on Computational Processing of the Portuguese Language*, pages 170–179.

Tsunenori Ishioka and Masayuki Kameda. 2006. Automated Japanese essay scoring system based on articles written by experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 233–240. Association for Computational Linguistics.

Benoit Lemaire and Philippe Dessus. 2003. A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, pages 305–320.

Ling Li and Hsuan-Tien Lin. 2007. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 NIPS Conference*. The MIT Press.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2021. Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2022. Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, pages 65–76.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica*, pages 276–82.

R. Mezher and Nazlia Omar. 2016. A hybrid method of syntactic feature and latent semantic analysis for automatic arabic essay scoring. *Journal of Applied Sciences*, pages 209–215.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, pages 238–243.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2021. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities.

Silvério Sirotheau, Eloi Favero, João Alves dos Santos, Simone Negrão, and Marco Nascimento. 2021. Avaliação automática de redações na língua portuguesa baseada na coleta de atributos e aprendizagem de máquina. In *Ciência da Computação: Tecnologias Emergentes em Computação*, volume 2, pages 56–68. Editora Científica Digital.

Leslie N. Smith. 2015. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186.

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to identify sentence parallelism in student essays. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 794–803.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, pages 403–417.

Bingjie Yang, Shengjie Zhao, Kenan Ye, and Rongqing Zhang. 2022. Distribution consistency penalty in the quadratic kappa loss for ordinal regression of imbalanced datasets. In *Proceedings of the Fifth International Conference on Computer Science and Artificial Intelligence*, pages 415–421.