ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355261259

Essay-BR: a Brazilian Corpus of Essays

Conference Paper · October 2021

DOI: 10.5753/dsw.2021.17414

CITATIONS
5

READS
113

3 authors:

Jeziel Marinho
Universidade Federal do Piauí
5 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Rafael Anchiêta
Instituto Federal de Educação, Ciência e Tecnologia do Piauí (IFPI)
32 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Raimundo Santos Moura
Universidade Federal do Piauí
38 PUBLICATIONS **198** CITATIONS
SEE PROFILE

All content following this page was uploaded by Rafael Anchiêta on 09 January 2023.

The user has requested enhancement of the downloaded file.

Essay-BR: a Brazilian Corpus of Essays

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

[1]Federal University of Piauí (UFPI) – Teresina, PI – Brazil

[2]Federal Institute of Piauí (IFPI) – Picos, PI – Brazil

{jezielcm, rsm}@ufpi.edu.br, 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. We create a large corpus with several essays written by Brazilian high school students on an online platform in order to bridge this gap. All of the essays are argumentative and were scored across five competences by experts. Moreover, we conducted an experiment on the created corpus and showed challenges posed by the Portuguese language. Our corpus is publicly available at* `https://github.com/rafaelanchieta/essay`.

1. Introduction

The Automated Essay Scoring (AES) area began with Page in 1996 [Page 1966] with 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 grade 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™, Criterion℠, IntelliMetric™, 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. 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 create a large corpus, namely Essay-BR, with essays written by Brazilian high school students through an online platform. 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.

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. To the best of our knowledge, this is the first publicly available corpus of essays for the Portuguese language, which meets the new ENEM evaluation criteria. We believe it will foster AES studies for that language, resulting in the development of alternative methods to grade an essay.

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 our conducted experiments. Finally, Section 6 concludes the paper, indicating future work.

2. Related Work

As before mentioned, there is no publicly available corpus of essays for the Brazilian Portuguese that meet the new ENEM assessment criteria. However, some efforts investigated AES for that language. Here, we briefly present them.

[Bazelato and Amorim 2013] crawled 429 graded essays from the *Educação UOL* Website 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[1].

[Amorim and Veloso 2017] developed an automatic essay scoring method for the Brazilian Portuguese language. For that, they collected $1,840$ graded essays from the *Educação UOL* Website. 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 the [Bazelato and Amorim 2013] work, the collected essays are very old and do not meet the ENEM exam criteria[2]. In a posterior work, [Amorim et al. 2018] analyzed the presence of biased ratings in the AES area.

[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

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

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

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 these works have used essays written in Brazilian Portuguese to evaluate their methods, the authors did not make corpora 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 detailed the Essay-BR corpus.

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 repurposed 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. Conform to the argumentative text genre and the proposed topic (prompt), to develop a text, using knowledge from different areas.
3. Select, relate, organize, and interpret data and arguments in defense of a point of view.
4. Usage of argumentative linguistic structures.
5. Elaborate 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 1. 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.

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.

Table 1. 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

4. Essay-BR Corpus

The Essay-BR corpus contains $4,570$ argumentative documents and 86 topics (prompts). They were collected from December 2015 to April 2020. 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 2 summarizes the Essay-BR corpus.

Table 2. Summary of the Essay-BR Corpus.

Details	Essay-BR
Text type	Argumentative
Writer's language level	BR students (high school)
Scoring	Holistic
Number of essays	$4,570$
Number of prompts	86
Number of competences	5
proficiency range	$[0 - 200]$
proficiency scores	$0, 40, 80, 120, 160, 200$
Score range	$[0 - 1,000]$

4.1. Construction of the Essay-BR corpus

To create the Essay-BR corpus, we developed a Web Scraper to extract essays from two public Websites: *Vestibular UOL*[3] and *Educação UOL*[4]. 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. We collected 798 essays and 43 prompts from *Educação UOL*, and $3,772$ essays and 276 prompts from *Vestibular UOL*. The difference in the number of essays is because of the latter Website receives up to forty essays per month, while the former receive up to twenty essays per month.

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

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

scoring strategy. Thus, we mapped the scores from Websites to ENEM scores, as shown in Figure 1.

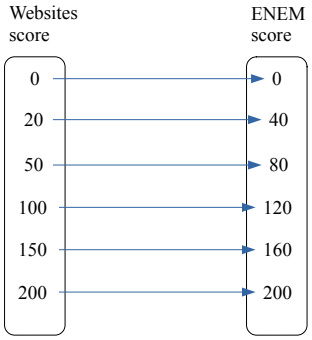

Figure 1. Mapping function from Website scores to ENEM scores.

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 3. 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 3. Example of the 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 corpus, as depicted in Tables 4 and 5, respectively. In Table 4, we can see that, on average, an essay has 4 paragraphs, and each paragraph has 2 sentences. Furthermore, the sentences are somewhat long, with an average of 30 tokens. In Table 5,

one can see that most essays are in the passive voice. This is because, in Portuguese, they 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, 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 4. Statistics of the essays.

Statistic	Mean	Std
Paragraph per essay	4.08	1.15
Sentence per essay	10.57	4.42
Sentence per paragraph	2.58	1.44
Token per essay	324.40	94.19
Token per paragraph	79.33	35.22
Token per sentence	30.66	17.68

Table 5. Linguistics features.

Feature	Mean
Passive voice	75%
Active voice	25%
lexical diversity	26%
lexical density	22%
Flesch score	45.98
Hapax legomenon	36.46

In order to organize the corpus, we divided it in proportions of 70%, 15%, and 15%, which corresponds to 3,198, 686, and 686 essays for training, development, and testing, respectively. Aiming to choose essays with a fair distribution of scores for each split, we computed the distribution of the total score of the essays, as depicted in Figure 2.

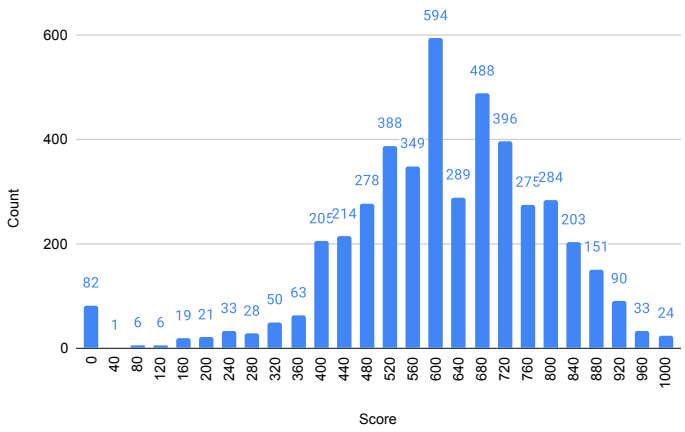

Figure 2. Distribution of the total score.

The top 3 scores are 600, 680, and 720 corresponding to 13.00%, 10.68%, and 8.67% of the 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 6. From this table, all of the essays received 120 as the higher score, showing that, in general, the students have medium dominance in all fields of competence.

In the following subsection, we analyzed the training, development, and testing sets of the Essay-BR corpus.

Table 6. Distribution score for each competence.

Competence	Scores					
	0	40	80	120	160	200
C1	97	24	359	**2,630**	1,338	122
C2	109	79	689	**1,711**	1,705	277
C3	122	146	1,206	**2,130**	827	139
C4	134	61	590	**2,000**	1,241	544
C5	302	276	1,023	**1,732**	815	422

4.2. Analysis of the Essay-BR 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 3 presents the score distribution for the training, development, and testing sets, respectively.

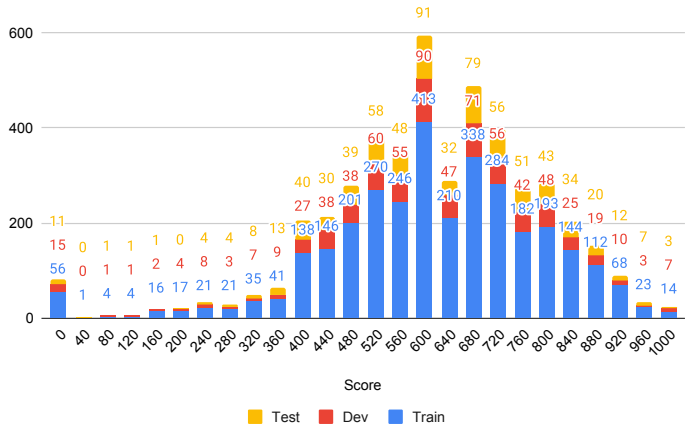

Figure 3. Training, development, and testing sets of the Essay-BR corpus.

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 2, 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 6 with the got results of each split in Table 7, 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 what follows, we present the experiment and obtained results.

Table 7. Statistics for each split of the corpus.

Split	Statistic	Mean	Standard deviation
Train	Paragraph per essay	4.08	1.08
	Sentence per essay	10.60	4.45
	Sentence per paragraph	2.59	1.44
	Token per essay	325.01	94.38
	Token per paragraph	79.52	35.00
	Token per sentence	30.64	17.70
Dev	Paragraph per essay	4.15	1.60
	Sentence per essay	10.57	4.63
	Sentence per paragraph	2.54	1.43
	Token per essay	323.52	95.69
	Token per paragraph	77.84	35.65
	Token per sentence	30.60	17.75
Test	Paragraph per essay	4.03	0.93
	Sentence per essay	10.45	4.10
	Sentence per paragraph	2.59	1.43
	Token per essay	322.48	91.75
	Token per paragraph	79.95	35.78
	Token per sentence	30.83	17.53

5. Experiments and Results

We carried out an experiment on the Essay-BR 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 the Essay-BR corpus 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 8 shows the QWK metric results, while Table 9 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.

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 a corpus with more than $50,000$ essays, whereas our corpus has 4.570 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 8. Quadratic Weighted Kappa on the test set.

Model	C1	C2	C3	C4	C5	Total
[Amorim and Veloso 2017]	0.35	0.44	0.39	0.37	0.34	0.47
[Fonseca et al. 2018]	0.42	0.46	0.40	0.45	0.36	0.51

Table 9. Rooted Mean Squared Error on the test set.

Model	C1	C2	C3	C4	C5	Total
[Amorim and Veloso 2017]	34.82	36.14	40.27	42.23	49.02	163.92
[Fonseca et al. 2018]	34.16	35.76	40.02	40.48	48.28	159.44

6. Final remarks

In this paper, we presented a large corpus of essays written by Brazilian high school students that were graded by experts following the evaluation criteria of the ENEM exam. This is the first publicly available corpus for that language. At this time, it has $4,570$ essays and 86 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 10. 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 modeling in different levels of the text. Modeling these challenging dimensions may require understanding the essay content and ex-

Table 10. 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

ploring the semantic and discourse levels of knowledge. Thus, there exist several possible applications that the Essay-BR may be useful.

As future work, besides increasing the corpus, which is already in process, we intend to provide the essay corrections, aiming to develop machine learning models to learn from the corrections.

Acknowledgments

The authors are grateful to IFMA for supporting this work.

References

Amorim, E., Cançado, M., and Veloso, A. (2018). 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)*, pages 229–237, New Orleans, Louisiana. Association for Computational Linguistics.

Amorim, E. and Veloso, A. (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, Valencia, Spain. Association for Computational Linguistics.

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

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

Beigman Klebanov, B. and Madnani, N. (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, Online. Association for Computational Linguistics.

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

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

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau,

D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Shermis, M. D. and Barrera, F. D. (2002). Exit assessments: Evaluating writing ability through automated essay scoring. page 31.

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

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

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

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

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

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

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). 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*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

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

View publication stats