CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Automated Formal Register Scoring of Student Narrative Essays Written in Portuguese

Moésio Wenceslau da Silva Filho[1,3], André C. A. Nascimento[1], Péricles Miranda[1], Luiz Rodrigues[3], Thiago Cordeiro[3], Seiji Isotani[4,5], Ig Ibert Bittencourt[3,5], Rafael Ferreira Mello[1,2,3]

[1]Universidade Federal Rural de Pernambuco

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)

[3]NEES, Universidade Federal de Alagoas

[4]Universidade de São Paulo

[5]Harvard Graduate School of Education

{moesio.wenceslau,rafael.mello}@ufrpe.br

Abstract. *Automated essay scoring (AES) is the task of automatically assigning scores (i.e., grades) to written texts. Although AES has been widely studied in the literature (e.g., informational and argumentative essays), specific types of texts still need more attention. Narrative essays are characterized by texts describing personal experiences and stories, either real or fictional. In this work, we describe a study on scoring student essays written in Portuguese under the aspect of Formal Register, which evaluates aspects related to the use of Brazilian Portuguese formal grammar and proficiency. The dataset created in this study provides a rich corpus of narrative essays produced in the context of a motivational situation, with a diverse set of language proficiency levels annotated by two professional graders. Different machine learning algorithms were evaluated using a diverse set of handcrafted linguistic features, and their results were compared against manual scores by the two human annotators. The results of the proposed analysis demonstrated that the AES model proposed achieved an equivalent agreement to that of the two human annotators.*

1. Introduction

Automated essay scoring (AES) aims to assign scores to students' essays automatically [Ferreira-Mello et al. 2019, Crossley 2020, Uto et al. 2020, de Lima et al. 2023]. However, most recent works have been focused on informational and argumentative essays, but other types of textual genres are gradually receiving more attention. Narrative essays are essential at the elementary and middle school levels and can be described as texts where the students share personal experiences and stories that can be either real or fictional [Somasundaran et al. 2018]. Narratives are used in numerous capacities at school instruction and assessment, as is the case in Brazilian public schools [Coelho 2020].

There are currently three main strategies to produce AES systems based on machine learning, according to the types of features used: (a) hand-engineered features, usually incorporating linguistic aspects and Natural Language Processing (NLP) frameworks, such as Coh-Metrix [Graesser et al. 2004, Cavalcanti et al. 2021b] to produce a

numeric representation of the essay; (b) raw-text based statistical models, often based on neural networks, that aim to map words and the essay to a high dimensional dense vector space (i.e., an embedding) [Ke and Ng 2019, Iqbal et al. 2023]; and (c) hybrid features, which combine the previous approaches, also usually in a neural network, to produce a more complex embedding [Uto et al. 2020]. Based on these representations, a supervised machine learning (ML) model can be trained to score the text according to specific criteria.

One of the challenges in developing such research in the context of narrative essays is the need for a dataset with annotated/scored essays, especially in Brazilian Portuguese. This perspective corroborates the findings from the literature review from [Bai and Stede 2022, Oliveira et al. 2023], which revealed a strong tendency of AES research in the English language but a lack of studies on other languages. Therefore, we describe a detailed pipeline to produce the first study on scoring student narrative essays in Portuguese under the **Formal Register** competency, which focuses on evaluating language proficiency regarding vocabulary usage, orthography, syntactic and lexical consistency. Specifically, a dataset of 327 human-annotated narrative essays (after quality assessments and pre-processing steps) was considered in this study.

Therefore, the primary objective of our study was to explore the feasibility of using machine learning techniques for estimating human ratings of formal writing scores in essays. To accomplish this goal, we extracted various linguistic features proposed in previous theoretical and empirical studies [Llach 2011, Hládek et al. 2020, Gimenes et al. 2015] from student essays written in Portuguese. We developed six models using these features and compared their performance. Specifically, we formulated the following research question to guide our investigation at this stage:

- *How accurate are machine learning algorithms in estimating scores generated by humans for formal register essays written in Portuguese?*

2. Related Works

Significant work has been done in the AES development [Fonseca et al. 2018, Ferreira-Mello et al. 2019, Ahadi et al. 2022], with some already being used in practice [Ke and Ng 2019, Shin and Gierl 2021]. However, previous studies have not focused on scoring narrative essays [Jones et al. 2019, Batista et al. 2022].

In [Somasundaran et al. 2018], two human annotators produced a dataset, scoring 942 essays for narrative-relevant aspects. The scoring rubric was done along three narrative traits: *Organization*, *Development*, and *Conventions*. Inter-annotator agreement over 344 doubly annotated essays was reported for each trait/sub-trait, using the Quadratic Weighted Kappa (QWK) metric, achieving a combined QWK of 0.76. The authors explored the methodology, demonstrating that it could reliably score development and organization traits in narratives when assessed by humans. The authors also explored several narrative-specific features for AES in narrative essays (e.g., linguistic and graph-based features), combined with four supervised learning algorithms (Linear Regression, Support Vector Regression, Random Forests, and Elastic Net).

Other works have explored different strategies to overcome the limitations of conventional AES, which typically relies on handcrafted features. In such a scenario, Deep

Neural Networks (DNN) are used to combine content-based representations (i.e., language models) with handcrafted or linguistic-based features to produce an AES [Crossley 2020, Shin and Gierl 2021, Uto et al. 2020], aiming to ease the feature engineering step. In [Uto et al. 2020], a diverse range of length and word-based, syntactic, and readability features are merged with a DNN-based distributed essay representation vector. The results demonstrated that incorporating handcrafted features improved the results over pure DNN models. However, such DNN-based approaches still lack interpretability, especially when used in an educational context [Ahadi et al. 2022].

Among the several criteria that can be used to evaluate a narrative text, assessing the level of students' formal writing (e.g., vocabulary usage, orthography, synctactic and lexical consistency) is critical as it influences other criteria [Llach 2011]. Previous work has already been proposed on spelling correction algorithms [Etoori et al. 2018, Hládek et al. 2020], and more specifically in Portuguese [Gimenes et al. 2015]. However, such algorithms focus only on detecting and possibly correcting errors and do not produce an overall formal register score. Notably, recent research has highlighted the need for exploring AES in languages other than English to understand its potential [Bai and Stede 2022]. The rationale is that once each language has its specific features, understanding how those affect AES's performance is prominent in shedding light on language-independent models.

3. Method

3.1. Study context and Dataset

This study utilized a comprehensive scoring rubric to evaluate the narrative construction proficiency of students' essays in late elementary school. The rubric offered clear guidelines for teachers to assess four key competencies: (i) **Formal Register**, (ii) **Thematic Coherence**, (iii) **Textual Typology**, and (iv) **Cohesion**. Each dimension was evaluated on a scale of 1 to 5 integer score points, with higher scores indicating better text quality and language proficiency.

In this study, we targeted the automatic scoring of the Formal Register, which evaluates the proper use of Brazilian Portuguese grammar and language proficiency. Specifically, we analyzed aspects such as vocabulary adequacy, including the usage of non-contextualized or unnecessary words, as well as the presence of oral language elements. Besides, we assessed lexical and syntactical features, such as appropriate verbal conjugation, nominal/verb agreement, and nominal/verb regency, along with correct spelling, punctuation, and segmentation of words within sentences, including deviations of hyposegmentation (i.e., union of unrelated words) and hypersegmentation (i.e., disconnecting components of a single word). Our comprehensive analysis of these key components allowed for a more in-depth understanding of the students' writing skills and provided a valuable resource for improving their writing proficiency.

The dataset developed consisted of narrative essays written by students aged 12 to 15 years. The task given to the students was to describe a fictional experience in a narrative text based on an initial triggering situation. An example of one essay and the suggested triggering situation to guide the student activity is provided in Table 1.

The final dataset comprised 327 narrative essays, with summary statistics presented in Table 2. On average, the essays comprised 135.1 words (excluding punctuation

Table 1. Example of an English translated essay and a triggering situation.

Essay	Triggering Situation
It was raining a lot that day, with very loud thunder coming from the sky. And after the rain passed I found in the backyard of my house a very giant shiny stone I could not believe what I was seeing, I was delighted with it took it and went inside without believing and kept in a safe, without knowing if it was really real	It rained a lot that day, with very loud thunder coming from the sky. And after the rain passed, I found in the backyard of my house a very shiny stone.

and white spaces) and 12.1 sentences. During the grading process, two human annotators independently evaluated the essays. To assess the discordance, we established a committee of three human annotators, which included the initial two and a third annotator with more experience in evaluating this type of text. This committee was responsible for determining the final score for each text.

Table 2. Summary statistics for the written essays in the dataset.

No. of Essays	327
Average Words	135.1 ± 68.4
Average Sentences	12.1 ± 10.1
Total Words	44176
Total Sentences	3957

3.2. Feature Extraction

The feature extraction process produced 8 independent features based on the aspects mentioned in Section 3.1, which were later combined to form the input feature vector. All features are real-valued numbers in the interval $[0, 1]$, indicating how "good" the essay was for that specific aspect. The features followed the same strategy as in [Amorim and Veloso 2017, Mello et al. 2021], in which they were derived from the number of occurrences of a characteristic in the input text. Furthermore, based on the need for further explorations of AES in languages other than English [Bai and Stede 2022], we chose those features based on expert guidelines for assessing formal register in the Portuguese language. Specifically, we defined and calculated this study's features as follows:

1. **Verb Agreement Score**: verb agreement, or subject-verb agreement, in Brazilian Portuguese dictates that verb and subject must agree in *number* and *gender*. This score was calculated by finding all instances of a subject followed by a verb and checking whether they agree in number and gender. If they agreed in both, this instance counted as a *hit*. Otherwise, it counted as an *error*. The feature was then calculated by $\frac{\text{hits}}{\text{hits} + \text{errors}}$ (when no instances were found, a default value of 1 was returned). This feature was implemented using spaCy [Honnibal et al. 2020].
2. **Nominal Agreement Score**: a nominal agreement in Brazilian Portuguese dictates that the elements (article, adjective, pronoun) modifying a noun must agree

in *number* and *gender* with it. This score was calculated by checking a predefined set of nominal agreement rules and storing whether they were correctly used in the text. Then, the feature was given by $\frac{\#\,\text{rules\_not\_followed}}{\#\,\text{rules}}$. It was implemented using the rules of CoGrOO [Kinoshita et al. 2006] and LanguageTool[1].

3. **Verb Conjugation Score**: In Brazilian Portuguese, verb conjugation score refers to the concept of verb regency, which limits the set of terms that can follow a given verb. To calculate this score, we first identified whether a token was a verb and then checked its associated lemma against a manually compiled dictionary of regencies based on Portuguese grammar. If the term matched a dictionary entry, it was counted as a hit. Otherwise, it was considered an error. This feature was calculated and implemented similarly to that of the verb agreement score.
4. **Nominal Conjugation Score**: In Brazilian Portuguese, nominal conjugation, or nominal government, restricts the set of terms that might follow a noun. This score was calculated as follows: given a token, check if it was a noun, then verify the possible terms associated with its lemma on a dictionary of regencies, built manually according to Portuguese grammar. If the term was in the dictionary, it was a hit; otherwise, it was an error. We calculated the Nominal Regency Score following the approach to extract the verb agreement score.
5. **Verb Conjugation Score**: Verb conjugation in Brazilian Portuguese defines the proper derived forms of a verb. This score was calculated by checking the verb against its possible forms. The feature was calculated analogous to the verb agreement score and was implemented using spaCy [Honnibal et al. 2020].
6. **Orthography Score**: We applied the LanguageTool[1] framework, considering the following list of errors: "Possible spelling error found", "Foreign words with diacritics", "Use of apostrophes for plural words", "Irregular feminine spelling", "Misspelling: Internet Abbreviations", "Easily Mistaken Rare Words", and "Rare words: Capitalization of geographic names". Each of these errors was checked for every token in the text. The feature was then calculated by $\frac{\#\,\text{errors}}{\#\,\text{tokens}}$ (when there were no tokens, a default value of 1 was returned).
7. **Hypersegmentation Score**: This feature splits a word/sentence into two or more tokens when it should be kept as one. It checked the existence of each token in a dictionary. If a word was not found, then we verified if the concatenation with its successor token (or a distance one variation of it) existed. In this case, it counted as a hyper-segmentation error; otherwise, it was a different type of error and was not counted toward this score. The feature was then calculated analogously to the orthography score.
8. **Hyposegmentation Score**: This feature is designed to identify when a student combines two words that should be separate. It utilizes Norvig's Segmentation Expansion algorithm [Segaran and Hammerbacher 2009], along with a custom word frequency, to detect hyposegmentation errors for every token in the input text. The feature was calculated similarly to that of the orthography score.

3.3. Model Selection and Evaluation

To address the proposed research question, we assessed the performance of several conventional machine learning algorithms using the features described in Section 3.2. We

[1] `https://github.com/languagetool-org/languagetool`

applied traditional algorithms (Decision Tree, Random Forest, and SVM) and state-of-the-art decision tree approaches (Extra-Tree, AdaBoost, and XGBoost). These classifiers were selected based on their performance in previous studies in natural language processing applications for education [Nau et al. 2020, Mello et al. 2021, Ahadi et al. 2022]. It is important to highlight that we modeled the problem as a supervised classification problem with five classes, i.e., grades 1-5 were treated as labels.

All experiments were performed with the scikit-learn toolkit. Due to resource limitations, we could not conduct hyper-parameter tuning within the cross-validation model selection process. Therefore, we recommend that future research investigate alternative algorithms and explore hyper-parameter tuning to assess the extent to which the predictive performance reported in this study holds up and can potentially be improved.

We evaluated the selected algorithms with the commonly used classification measures for educational data mining and learning analytics [Hossin and Sulaiman 2015]: precision, recall, F_1-score, and Cohen's κ. We performed a two-step process to assess the quality of the proposed approach. Initially, we considered the dataset described in section 3.1 and applied a stratified K-fold cross-validation (CV) with $K = 5$. We recorded the weighted precision, recall, F_1-score, and Cohen's κ for each CV iteration.

Secondly, we performed an agreement analysis between the algorithm with the best results and the human annotators in the original dataset of 327 essays described in section 3.1. In this case, we adopted the McNemar's test [Raschka 2018], with $\alpha = 0.05$, to check whether there existed a significant difference between the model we trained and the human annotators.

4. Results

The results of the stratified cross-validation, presented in Table 3, include the average weighted precision, recall, F_1-score, and Kappa metrics. Among the algorithms evaluated, the Extra-Tree ensemble demonstrated the best performance across all measures, while the SVM performed the poorest. The Random Forest algorithm produced results comparable to those of the Extra-Tree ensemble. XGBoost and Adaboost achieved Kappa scores of 0.256 and 0.253, respectively.

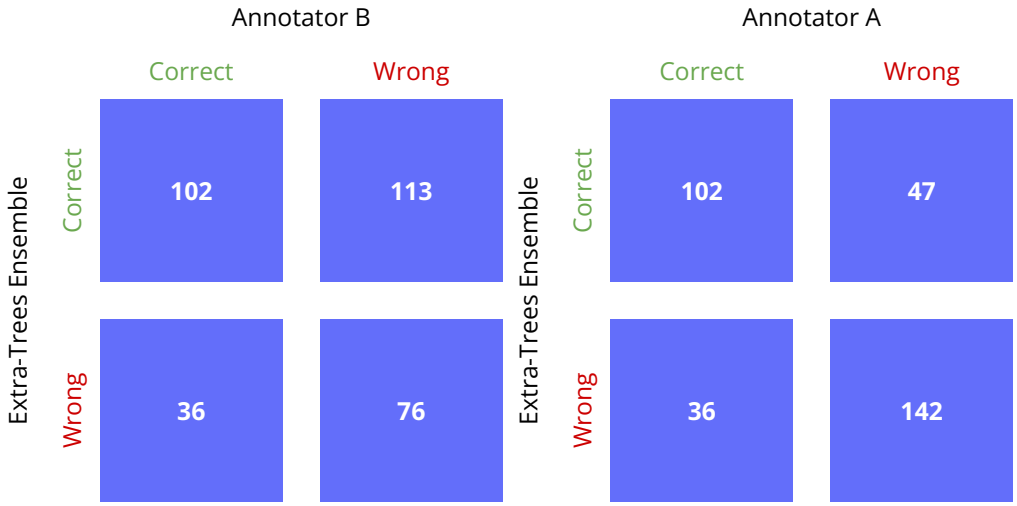
Table 3. Average weighted precision, recall and F_1-score over 5-fold cross-validation.

Algorithm	Precision	Recall	F_1-score	Kappa
SVM	0.411 ($\pm$0.111)	0.468 ($\pm$0.053)	0.412 ($\pm$0.075)	0.171 ($\pm$0.085)
Decision Tree	0.460 ($\pm$0.086)	0.453 ($\pm$0.068)	0.450 ($\pm$0.072)	0.236 ($\pm$0.104)
Random Forest	0.548 ($\pm$0.066)	0.560 ($\pm$0.062)	0.539 ($\pm$0.061)	0.354 ($\pm$0.080)
XT Ensemble	**0.557** ($\pm$0.047)	**0.566** ($\pm$0.045)	**0.546** ($\pm$0.046)	**0.367** ($\pm$0.052)
Adaboost	0.463 ($\pm$0.056)	0.464 ($\pm$0.047)	0.439 ($\pm$0.059)	0.253 ($\pm$0.071)
Xgboost	0.467 ($\pm$0.057)	0.485 ($\pm$0.050)	0.468 ($\pm$0.054)	0.256 ($\pm$0.068)

Since the Extra-Tree ensemble achieved the highest performance in the previous analysis, it was chosen to assess the agreement with the human annotators. Figure 1

presents the contingency tables of McNemar's test for both annotators, showing the number of correct and incorrect predictions/annotations made by the algorithm and one human annotator compared to the other. For example, the first matrix in Figure 1 compares the model's predictions/annotations with annotator B's for the categories included by annotator A. It reveals that the model correctly predicted 215 instances, while the annotators agreed on only 138. Moreover, the right side of Figure 1 shows that the model agreed with more instances (148) than the human annotators agreed on themselves (138).

In addition to the results in Figure 1, McNemar's test revealed that the difference between the model and annotator B regarding annotator A's annotations was statistically significant (p-value $\approx 4.78\text{e}{-10}$). This implies that the model outperformed annotator B in this case, meaning that the model had a higher agreement with annotator A than the agreement between the two human annotators A and B. Conversely, in the second case (model vs. annotator B concerning annotator A), the difference was not statistically significant (p-value ≈ 0.27).



(a) Against annotator A **(b)** Against annotator B

Figure 1. McNemar Contingency Tables for the Extra-Trees Ensemble and annotators A and B.

5. Discussion

The results for the automatic classification of the categories for formal register in narrative essays written in Brazilian Portuguese demonstrated that the proposed approach, based on hand-engineering features, reached performance compared to human annotators. In the best-case scenario, the Extra-Trees model reached 0.526 Cohen's κ, representing a moderate level agreement rate [Landis and Koch 1977]. It is important to highlight that XGboost did not obtain the best result for this task, which we expected as it usually performs better when traditional machine learning algorithms are applied to AES [Ramesh and Sanampudi 2022]. We hypothesized that it happened because XGboost is more robust for larger datasets[Chen and Guestrin 2016].

The current study highlighted the difficulty of annotating essays in Portuguese, this challenge arises due to the numerous variations of similar grammar rules in the Por-

tuguese language, making the analysis more subjective [Amorim and Veloso 2017, Marcilese et al. 2019]. Consequently, this can impact the performance of machine learning models. Thus, further research should explore this direction to improve the accuracy of these models. Additionally, Marcilese et al. [Marcilese et al. 2019] underscore that the Brazilian Portuguese language has extensive nonstandard linguistic variations, which coexist with a standardized variety defined by the language's rules. Their findings also suggest that formal schooling plays a crucial role in children mastering the rules of the formal Portuguese language. In this context, our study provides automated means to assist in evaluating students' understanding of the formal language.

There are three key implications of our study for research and practice. Firstly, we created a new dataset of narrative essays written by students aged 12-15 years from Brazilian public schools. To the best of our knowledge, this is the first dataset of essay scoring for narrative texts in Portuguese. The final version of the dataset, which will be released by the end of 2023, will contain 1300 essays graded according to the four criteria aforementioned in 3.1. Secondly, our results show that the proposed classifier, developed using hand-crafted features and the Extra-Tree ensemble, achieved results comparable to the human annotators. Therefore, it could serve as the foundation for creating an AI-driven system to support student practices, potentially reducing the time instructors need to review and score each essay. Thirdly, our study is part of a national-level project entitled "**Plataforma Adaptativa de Avaliação e Diagnóstico Pedagógico de Textos**", aimed at addressing the learning gap among students after the pandemic. As such, the developed model will be applied in a real-world scenario in the near future.

6. Limitations and directions for future research

We acknowledge the following limitations of the study. First, although we explored a wide range of traditional machine learning algorithms, we have not explored DNN-based architectures or neural language models. Such models would allow a more comprehensive evaluation, including punctuation and semantic aspects. Furthermore, recent works [Uto et al. 2020] demonstrated that combining language models with handcrafted features can improve performance. In future works, we intend to explore hybrid methods proposed in previous works [Uto et al. 2020, Yuan et al. 2020].

Secondly, it is worth mentioning that the model did not include punctuation or semantic features due to their high computational cost [Lima et al. 2022]. Future work should aim to incorporate these features into the final model to further improve its performance.

Moreover, an in-depth analysis of the most relevant features selected by the models may lead to interesting discoveries, which were out of the scope of this paper. Furthermore, a thorough investigation of techniques for dealing with the imbalanced and possibly conflicting nature, due to having two annotators, of the dataset might also contribute to a better understanding of AES. Finally, this study did not involve the analysis of the impact of the use of AES on a real-world scenario. A future extension to this work would be the assessment of instructors' and students' satisfaction as well as the impact of the use of such methods on the development of students' writing skills [Cavalcanti et al. 2021a].

References

Ahadi, A., Singh, A., Bower, M., and Garrett, M. (2022). Text mining in education—a bibliometrics-based systematic review. *Education Sciences*, 12(3):210.

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.

Bai, X. and Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, pages 1–39.

Batista, H. H., Barbosa, G. A., Miranda, P., Santos, J., Isotani, S., Cordeiro, T., Bittencourt, I. I., and Mello, R. F. (2022). Detecção automática de clímax em produções de textos narrativos. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 932–943. SBC.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., and Mello, R. F. (2021a). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.

Cavalcanti, A. P., Mello, R. F., Miranda, P., Nascimento, A., and Freitas, F. (2021b). Utilização de recursos linguísticos para classificação automática de mensagens de feedback. In *Anais do XXXII Simpósio Brasileiro de Informática na Educaçao*, pages 861–872. SBC.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Coelho, R. (2020). Teaching writing in brazilian public high schools. *Reading and Writing*, 33(6):1477–1529.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

de Lima, T. B., da Silva, I. L. A., Freitas, E. L. S. X., and Mello, R. F. (2023). Avaliação automática de redação: Uma revisão sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.

Etoori, P., Chinnakotla, M., and Mamidi, R. (2018). Automatic spelling correction for resource-scarce languages using deep learning. In *Proceedings of ACL 2018, Student Research Workshop*, pages 146–152, Melbourne, Australia. Association for Computational Linguistics.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Com-*

putational Processing of the Portuguese Language, pages 170–179, Cham. Springer International Publishing.

Gimenes, P. A., Roman, N. T., and Carvalho, A. M. (2015). Spelling error patterns in brazilian portuguese. *Computational Linguistics*, 41(1):175–183.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Hládek, D., Staš, J., and Pleva, M. (2020). Survey of automatic spelling correction. *Electronics*, 9(10).

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Iqbal, S., Rakovic, M., Chen, G., Li, T., Ferreira Mello, R., Fan, Y., Fiorentino, G., Radi Aljohani, N., and Gasevic, D. (2023). Towards automated analysis of rhetorical categories in students essay writings using bloom's taxonomy. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 418–429.

Jones, S., Fox, C., Gillam, S., and Gillam, R. B. (2019). An exploration of automated narrative analysis via machine learning. *Plos one*, 14(10):e0224634.

Ke, Z. and Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Kinoshita, J., Salvador, L. d. N., and de Menezes, C. E. D. (2006). CoGrOO: a Brazilian-Portuguese grammar checker based on the CETENFOLHA corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*.

Lima, T. B. D., Miranda, P., Mello, R. F., Wenceslau, M., Bittencourt, I. I., Cordeiro, T. D., and José, J. (2022). Sequence labeling algorithms for punctuation restoration in brazilian portuguese texts. In Xavier-Junior, J. C. and Rios, R. A., editors, *Intelligent Systems*, pages 616–630, Cham. Springer International Publishing.

Llach, M. d. P. A. (2011). *Lexical errors and accuracy in foreign language writing*. Multilingual Matters.

Marcilese, M., Name, C., Augusto, M., Molina, D., and Armando, R. (2019). Mother-tongue education, linguistic variation and language processing. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, 72(3):17–40.

Mello, R. F., Fiorentino, G., Miranda, P., Oliveira, H., Raković, M., and Gašević, D. (2021). Towards automatic content analysis of rhetorical structure in brazilian college

entrance essays. In *International Conference on Artificial Intelligence in Education*, pages 162–167. Springer.

Nau, J., Dazzi, R. L., Filho, A. H., and Fernandes, A. (2020). Processamento do discurso em textos dissertativos-argumentativos: Uma abordagem baseada em mineração de argumentos e aprendizado supervisionado de máquina. In *Anais do XLVII Seminário Integrado de Software e Hardware*, pages 48–59, Porto Alegre, RS, Brasil. SBC.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

Ramesh, D. and Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808.

Segaran, T. and Hammerbacher, J. (2009). *Beautiful data*. O'Reilly Media, Sebastopol, CA.

Shin, J. and Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, 38(2):247–272.

Somasundaran, S., Flor, M., Chodorow, M., Molloy, H., Gyawali, B., and McCulla, L. (2018). Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Uto, M., Xie, Y., and Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuan, Z., Jiang, Y., Li, J., and Huang, H. (2020). Hybrid-dnns: Hybrid deep neural networks for mixed inputs. *arXiv preprint arXiv:2005.08419*.