# Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays

Rafael Ferreira Mello[1(✉)], Giuseppe Fiorentino[1], Péricles Miranda[1], Hilário Oliveira[2], Mladen Raković[3], and Dragan Gašević[3]

[1] Department of computing, Universidade Federal Rural de Pernambuco, Recife, Brazil
{rafael.mello,pericles.miranda}@ufrpe.br
[2] Instituto Federal do Espírito Santo, Vitória, Brazil
[3] Centre for Learning Analytics, Faculty of Information Technology, Monash University, Melbourne, Australia
{mladen.rakovic,dragan.gasevic}@monash.edu

**Abstract.** Essay scorers manually look for the presence of required rhetorical categories to evaluate coherence, which is a time-consuming task. Several attempts in the literature have been reported to automate the identification of rhetorical categories in essays with machine learning. However, existing machine learning algorithms are mostly trained on content features which can lead to over-fitting and hindering model generalizability. Thus, this paper proposed a set of content-independent features to identify rhetorical categories. The best performing classifier, XGBoost, achieved performance comparable to human annotation and outperformed previous models.

**Keywords:** Essay analysis · Content analytics · Rhetoric structure

## 1 Introduction

Essays are short literary compositions that reflect an author's perspective on a particular topic [30]. At Brazilian universities, the essay writing exam is among the key components in the admission process. In 2020, 5.8 million students had applied for university admission and took essay writing exams[1]; this increased the challenge to provide a good quality assessment for every essay submitted.

Prospective students are required to write a dissertative-argumentative essay that needs to be coherent and cohesive and written in Portuguese using a formal academic style. Students are required to compose sentences with different rhetorical functions (e.g., thesis, argument) and connect them into a coherent essay with an introduction, argumentation, and conclusion [16]. Thus, automatic

[1] https://bit.ly/36LivBB.


I. Roll et al. (Eds.): AIED 2021, LNAI 12749, pp. 162–167, 2021.
https://doi.org/10.1007/978-3-030-78270-2_29

identification of rhetorical categories in Brazilian entrance essays can improve the essay scoring process's efficiency during the university admission process.

To date, researchers have proposed several computational solutions to automatic identification of rhetorical components in student essays, e.g., [4,24,25]; of specific relevance is the work by dos Santos et al. [26] who propose an approach for the context of entrance essays at Brazilian universities. To this end, researchers have developed classification models based on machine learning (ML) and natural language processing (NLP) methods. While those classifiers demonstrated empirical validity and attractive classification performance (73%–93%), they typically involved some form of content features, i.e., features determined by the vocabulary in an essay (e.g., key terms, noun phrases). Due to differences in vocabulary use among students, however, reliance on content features can diminish classification performance, leading to over-fitting and hindering model robustness and generalizability [18].

In this study, we looked at the automatic identification of rhetorical components in student essays a step forward, as we developed a supervised machine learning model that relies upon content-independent features. To accomplish this goal, we examined a set of features derived from the two empirically validated linguistic tools, the Linguistic Inquiry Word Count (LIWC, [28]) and Coh-Metrix [15], and also computed order and relational features. The features in our study measured text cohesion, readability, and semantic relations and are not affected by a student's vocabulary use. Further, we developed five machine learning classification models to identify rhetorical categories in the essays.

## 2 Method

### 2.1 Dataset

The structure of the Brazilian essay writing exam was detailed in [26]. This framework represents characteristics of genre observed in essays written by students on entrance exams. Since the students are required to write an opinion paper on an assigned topic, it is common for their write-ups to start with the title (class **s0**). Introductory sentences that follow contextualize and present the subject addressed in the text.

These sentences are classified as Theme (**t1**). Thereafter, students are required to present their viewpoints regarding the essay topic. These sentences are categorized as Thesis (**t2**). After the introduction, students must present factual and logically valid arguments to support their thesis. These sentences are classified as (**s2**). The argumentation section is generally most elaborate in the essay, as students need to discuss their viewpoints. Finally, students conclude their essay. Conclusions commonly involve sentences summarising the initial thesis (classified as Background, **t3**) and sentences presenting final arguments (classified as Conclusion **s3**) that may or may not offer solutions to the problem discussed in the essay.

In this study, we used the dataset initially created by [26]. It encompasses 271 essays, divided into 2,562 sentences, written by candidates that took entrance

exams applying for Brazilian universities in 2014 and 2016. Three human annotators with background in computer science and linguistics coded each sentence in the essay corpus according to the following categories: Title, Theme, Thesis, Argumentation, Background, Conclusion, and Author. The value of Fleiss's $\kappa$ agreement between the coders reached 0.78. The disagreements were resolved by adopting the category the majority of coders elected for.

### 2.2 Feature Extraction

We examined the performance of predictive models that use features based on linguistic resources for rhetorical structure identification. Those features have been largely harnessed in other problems in educational research [7,13,23].

**LIWC Features:** The Linguistic Inquiry Word Count (LIWC) is a dictionary of measures indicative of different psychological processes (e.g., affective, cognitive, social, perceptual) [28]. In this study, we utilized the Portuguese version of LIWC proposed in 2019 [6]. The Portuguese version contains 73 categories of word counts that were used as features in this paper.

**Coh-Metrix Features:** Coh-Metrix is a computational linguistics tool that provides measures of linguistic complexity, text coherence, text readability, and lexical category [22]. It has been widely used in previous studies to analyze essay coherence and structure (e.g., [1,11,21]). The Portuguese version of Coh-Metrix used in this paper [5] has 98 different measures.

**Ordering Features:** In addition to the indicators implemented by LIWC and Coh-Metrix, different theories indicated that capturing the flow of the ideas in the document is essential to categorize text blocks into a rhetorical structure model [29]. Therefore, we also incorporated two features capturing the order of the sentences in the text: i) the position from the first sentence to the last; ii) the position of the sentences from the last to the first.

**Features Extracted from Adjacent Sentences:** The initial feature space used in this work had a total of 173 features. However, previous works [14,26] indicated that the use of sequence-based machine learning models could reach better results for this problem. We adopted the features extracted from the actual sentence and the previous and following sentences for each sentence to incorporate the notion of sequence into traditional machine learning algorithms. Thus, the final feature space in our analysis contained 519 features.

### 2.3 Model Selection and Evaluation

We trained several machine learning classifiers, including Random Forest, Gaussian kernel SVM, AdaBoost, XGBoost, and CRF. Random Forest and Gaussian kernel SVM were included based on the good performance of several previous analyses [12]. Moreover, AdaBoost and XGBoost decision tree algorithms have demonstrated better results when compared to Random Forest [8,9]. Finally, we also evaluated the Conditional Random Fields (CRF) algorithm's performance.

This algorithm is largely adopted for sequence labeling problems such as the analysis of rhetorical structures [27].

We used the same evaluation process performed in the previous work [26] to compare the classification results. To measure the performance of supervised machine learning algorithms, we adopted Cohen's $\kappa$ [10], a metrics commonly used in educational data mining and learning analytics [20,23], and precision, recall, and f-measure, which are widely used metrics in the field of machine learning [2]. We applied 10-fold stratified cross-validation to evaluate all the measures obtained.

## 3 Results

Table 1 shows the best results achieved by each algorithm using 10-fold cross-validation (as described in Sect. 2.3). The outcomes revealed that the XGBoost algorithm reached the best results in general for all metrics evaluated. The XGBoost outperformed, in terms of Cohen's $\kappa$, by 6.34% and 11.70% the CRF and Random Forrest classifiers, respectively. Adaboost and SVM achieved the worst results in this experiment.

**Table 1.** Results for the analysed algorithms in terms of precision, recall, F1-score, and Cohen's $\kappa$.

| Algorithm | Precision | Recall | F1-score | $\kappa$ |
|---|---|---|---|---|
| SVM | 0.56 | 0.58 | 0.50 | 0.42 |
| Random forest | 0.71 | 0.71 | 0.71 | 0.60 |
| AdaBoost | 0.65 | 0.55 | 0.59 | 0.45 |
| XGBoost | 0.73 | 0.75 | 0.73 | 0.67 |
| CRF | 0.70 | 0.72 | 0.71 | 0.63 |

## 4 Discussion and Practical Implications

The best performing classifier, XGBoost, achieved $\kappa$ of 0.67, the performance comparable to human annotation as discussed in [26]. Importantly, as a part of our modeling approach, we utilized the non-content features to improve the performance and generalizability of the classifier. As these features represent the structure (e.g., cohesiveness, legibility, semantic relationships, number of nouns and pronouns) of the text instead of the content itself [3,13], a considerably accurate classifier we developed based on those features promises robustness in predicting rhetorical categories across different writing styles and genres. Equally important, this approach to feature extraction reduces the total number of features in the model, decreasing the chances of over-fitting [18].

The study's practical implications include: (i) the classifier we developed may provide accurate automatic identification of rhetorical categories in entrance essays and reduce the time the assessors need to review and score each essay

manually; (ii) in the context of writing assignments in university courses, the automatic analysis of rhetorical structures could generate valuable information in creating formative feedback to guide essay revisions [17]; and (iii) the results provide a foundation for the development of learning analytics tools for instructors and students based on the rhetorical structure theory [19].

## References

1. Abba, K.A., Joshi, R.M., Ji, X.R.: Analyzing writing performance of l1, l2, and generation 1.5 community college students through coh-metrix. Written Lang. Literacy **22**(1), 67–94 (2019)
2. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining Text Data, pp. 163–222. Springer (2012). https://doi.org/10.1007/978-1-4614-3223-4_6
3. Barbosa, G., et al.: Towards automatic cross-language classification of cognitive presence in online discussions. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 605–614 (2020)
4. Burstein, J., Marcu, D., Knight, K.: Finding the write stuff: automatic identification of discourse structure in student essays. IEEE Intell. Syst. **18**(1), 32–39 (2003). https://doi.org/10.1109/MIS.2003.1179191
5. Camelo, R., Justino, S., de Mello, R.F.L.: Coh-metrix PT-BR: uma API web de análise textual para a educação. In: Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação, pp. 179–186. SBC (2020)D
6. Carvalho, F., Rodrigues, R.G., Santos, G., Cruz, P., Ferrari, L., Guedes, G.P.: Evaluating the Brazilian Portuguese version of the 2015 LIWC lexicon with sentiment analysis in social networks. In: Anais do VIII Brazilian Workshop on Social Network Analysis and Mining, pp. 24–34. SBC (2019)
7. Cavalcanti, A.P., et al.: How good is my feedback? A content analysis of written feedback. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 428–437 (2020)
8. Chan, J.C.W., Paelinckx, D.: Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing Environ. **112**(6), 2999–3011 (2008)
9. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
10. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
11. Crossley, S.A., McNamara, D.S.: Understanding expert ratings of essay quality: Coh-metrix analyses of first and second language writing. Int. J. Continuing Eng. Educ. Life Long Learn. **21**(2–3), 170–191 (2011)
12. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**(1), 3133–3181 (2014)
13. Ferreira, M., Rolim, V., Mello, R.F., Lins, R.D., Chen, G., Gašević, D.: Towards automatic content analysis of social presence in transcripts of online discussions. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 141–150 (2020)
14. Fiacco, J., Cotos, E., Rose, C.: Towards enabling feedback on rhetorical structure with neural sequence models. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 310–319 (2019)

15. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-metrix: providing multilevel analyses of text characteristics. Educ. Res. **40**(5), 223–234 (2011)
16. Haendchen Filho, A., do Prado, H.A., Ferneda, E., Nau, J.: An approach to evaluate adherence to the theme and the argumentative structure of essays. Proc. Comput. Sci. **126**, 788–797 (2018)
17. Jiang, S., Yang, K., Suvarna, C., Casula, P., Zhang, M., Rose, C.: Applying rhetorical structure theory to student essays for providing automated writing feedback. In: Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pp. 163–168 (2019)
18. Khalid, S., Khalil, T., Nasreen, S.: A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and Information Conference, pp. 372–378. IEEE (2014)
19. Kiesel, D., Riehmann, P., Wachsmuth, H., Stein, B., Froehlich, B.: Visual analysis of argumentation in essays. IEEE Trans. Visual. Comput. Graph. **27**, 1139–1148 (2020)
20. Kovanovic, V., Joksimovic, S., Gasevic, D., Hatala, M.: What is the source of social capital? The association between social network position and social presence in communities of inquiry. In: Workshop at Educational Data Mining Conference. EDM (2014)
21. Latifi, S., Gierl, M.: Automated scoring of junior and senior high essays using coh-metrix features: implications for large-scale language testing. Lang. Test. 0265532220929918 (2020)
22. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)
23. Neto, V., Rolim, V., Ferreira, R., Kovanović, V., Gašević, D., Lins, R.D., Lins, R.: Automated analysis of cognitive presence in online discussions written in Portuguese. In: European Conference on Technology Enhanced Learning, pp. 245–261. Springer (2018). https://doi.org/10.1007/978-3-319-98572-5_19
24. Nguyen, H., Litman, D.: Context-aware argumentative relation mining. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1127–1137 (2016)
25. Rakovic, M., Winne, P., Marzouk, Z., Chang, D.: Automatic identification of knowledge transforming content in argument essays developed from multiple sources. J. Comput. Assist. Learn
26. dos Santos, K.S., Soder, M., Marques, B.S.B., Feltrim, V.D.: Analyzing the rhetorical structure of opinion articles in the context of a Brazilian college entrance examination. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 3–12. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99722-3_1
27. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. In: Introduction to Statistical Relational Learning, vol. 2, pp. 93–128 (2006)
28. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010). https://doi.org/10.1177/0261927X09351676
29. Van Dijk, T.A.: Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Routledge (2019)
30. Zupanc, K., Bosnić, Z.: Automated essay evaluation with semantic analysis **120**(C), 118–132 (2017). https://doi.org/10.1016/j.knosys.2017.01.006