Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese

Rafael Ferreira Mello
Cesar School
Recife, PE, Brazil
Federal Rural University of Pernambuco
Recife, PE, Brazil
rflm@cesar.school

Giuseppe Fiorentino
Federal Rural University of Pernambuco
Recife, PE, Brazil
fiorentinogiuseppebcc@gmail.com

Péricles Miranda
Federal Rural University of Pernambuco
Recife, PE, Brazil
pericles.miranda@ufrpe.br

Hilário Oliveira
Instituto Federal do Espírito Santo
Brazil
hilariotomaz@gmail.com

Mladen Raković
Monash University
Melbourne, VC, Australia
mladen.rakovic@monash.edu

Dragan Gašević
Monash University
Melbourne, VC, Australia
University of Edinburgh
Edinburgh, U.K.
King Abdulaziz University
Jeddah, Saudi Arabia
Dragan.Gasevic@monash.edu

ABSTRACT

Brazilian universities have included essay writing assignments in the entrance examination procedure to select prospective students. The essay scorers manually look for the presence of required Rhetorical Structure Theory (RST) categories and evaluate essay coherence. However, identifying RST categories is a time-consuming task. The literature reported several attempts to automate the identification of RST categories in essays with machine learning. Still, previous studies have focused on using machine learning algorithms trained on content-dependent features that can diminish classification performance, leading to over-fitting and hindering model generalisability. Therefore, this paper proposes: (i) the analysis of state-of-the-art classifiers and content-independent features to the task of RST rhetorical moves; (ii) a new approach that considers the sequence of the text to extract features – i.e. sequential content-independent features; (iii) an empirical study about the generalisability of the machine learning models and sequential content-independent features for this context; (iv) the identification of the most predictive features for automated identification of RST categories in essays written in Portuguese. The best performing classifier, XGBoost, based on sequential content-independent features, outperformed the classifiers used in the literature and are based on traditional content-dependent features. The XGBoost classifier based on sequential content-independent features also reached promising accuracy when tested for generalisability.

CCS CONCEPTS

• **Information systems → Clustering and classification**; • **Applied computing → E-learning**.

KEYWORDS

Essay analysis, content analytics, context analysis, rhetoric structure, natural language processing.

ACM Reference Format:
Rafael Ferreira Mello, Giuseppe Fiorentino, Péricles Miranda, Hilário Oliveira, Mladen Raković, and Dragan Gašević. 2022. Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3506860.3506977

1 INTRODUCTION

The ability to effectively communicate information in a written form is one of the essential skills for professionals in the modern job market. Professionals are often required to compose argument-based texts that clearly present ideas and back those ideas up with relevant evidence. Therefore, educating students to write effective and well-supported argumentative essays is an important goal of contemporary education. For this reason, educators in many colleges and universities have included essay writing assignments in the entrance examination procedure to select prospective students [17]. At Brazilian universities, the essay writing exam is one of the key components in the admission process. In 2020, 5.8 million

LAK22, March 21–25, 2022, Online, USA

ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506977

students applied for university admission and took essay writing exams [1]. However, due to an enormous number of essays created each year in the admission process, educators at Brazilian institutions face a challenge to provide a timely, yet quality evaluation for every essay submitted. These include both summative and formative evaluations.

According to the guidelines proposed by the Brazilian National Institute of Educational Studies and Research[2], which is used in the admission process of several universities in Brazil, prospective students are required to write a dissertative-argumentative essay on an assigned social, scientific, cultural, or political topic. In this essay, students present a thesis on a topic and support it with plausible arguments. Moreover, the essay text needs to be coherent, 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, as defined by dissertative-argumentative genre [23].

In this context, the entrance essay score is determined by assessing how a written essay conforms to the expected rhetorical structure [52], i.e., the essay scorers look for the presence of required rhetorical categories and evaluate essay coherence. However, the manual identification of rhetorical categories and their relations in the essay is a time-consuming task, given the extensive number of candidates who take the university entrance exams every year. Moreover, inconsistencies can be introduced in the evaluation process since a single human assessor is usually required to analyse several essays [17]. Thus, automatic identification of rhetorical categories in Brazilian entrance essays can mitigate this problem towards more efficient essay evaluation during the university admission process.

Several studies have been conducted to propose and evaluate computational solutions to identifying rhetorical components in students' essays automatically [8, 29, 31, 46, 48]. In general, these studies use a typical classification approach that involves traditional machine learning algorithms in combination with content-dependent features based on the vocabulary of the essays in the training dataset (e.g., keywords, noun phrases). Furthermore, a few of these studies use sequence-based classifiers (i.e., CRF) [17] and sequence and process mining [30] to increase their performance by capturing the flow of rhetorical components in essays [37, 38].

However, due to differences in vocabulary usage among students and changes in topics addressed in essays in different contexts, reliance on content resources can decrease classification performance, which can lead to over-fitting, and thus decreasing the generalisability of the models [27]. To the best of our knowledge, no previous work has examined the effectiveness of the use of sequential features combined with traditional machine learning algorithms for automatic recognition of rhetorical components.

Therefore, this paper proposes a new approach for automated analysis of rhetorical structure in written essays that relies upon sequential content-independent features and a state-of-the-art supervised machine learning algorithm. More specifically, we investigate a set of content-independent features extracted from the two well-known linguistic tools, the Linguistic Inquiry Word Count (LIWC, [54]) and Coh-Metrix [22]. The adopted features measured text cohesion, readability, and semantic relations, which decreased the dependence on the content of the text. Moreover, we introduced a new method to extract features that account for sequences of sentences in essay text. This approach is widely adopted in sequence-based classifiers (i.e., CRF), promising to increase the classification performance and generalisability of traditional machine learning algorithms. Finally, we also provide a detailed analysis of the most predictive content-independent features for each rhetorical category, which could be used to support the provision of actionable feedback.

In order to assure the effectiveness of the proposed approach, we evaluated several supervised machine learning algorithms and feature sets used to detect rhetorical categories in Brazilian entrance essays. The study results demonstrated that the XGBoost classification model based on the proposed sequential content-independent feature set achieved the best performance in terms of accuracy and generalisability, compared to other classifiers we examined. Further, the feature analysis demonstrated that sequential content-independent features introduced in the current study were important predictors of rhetorical categories in the entrance college essays. Hence, our findings can inform the development of the systems that identify rhetorical categories automatically and that are more robust to essay content. This, in turn, could further boost reliability and facilitate the essay scoring process. Moreover, the results of feature analysis reported in this paper can potentially be used to inform the development of actionable, formative feedback for students.

2 BACKGROUND

2.1 Rhetorical Structure of Student Essays

To communicate their arguments effectively and fulfil the requirements of a writing task, writers structure their essays as sets of coherently connected rhetorical components, e.g., background information, claim and evidence. The extent to which an essay is well-structured (i.e., coherent) often determines the quality of the entire essay. For that reason, many researchers have studied the rhetorical structure of essays written in different contexts and for different purposes (e.g., [8, 17]. To this end, the Rhetorical Structure Theory (RST)[37, 38] has been one of the most widely utilised theoretical approaches. This theoretical framework has often underlined the development of analytical approaches with empirically documented benefits, e.g., automatic text summarisation [57] and discourse analysis of student writing [8, 17]. For a more comprehensive overview of the RST applications see [52].

According to RST, a writer structures a text as a sequence of interconnected Elementary Discourse Units (EDUs), often determined by genre requirements (e.g., argument essay). EDU, therefore, represents the minimal logical unit of meaning in the essay. In a grammatical sense, these units are often expressed as sentences [6]. RST further proposes that the text structure can be represented as a tree, where EDUs are nodes and rhetorical relations among them are edges. Over the decades of using RST for discourse analysis, researchers have defined many different types of rhetorical categories and proposed different relationships among them. These

[1] https://bit.ly/36LivBB
[2] https://www.gov.br/inep

categories and relationships have usually been imposed by genre requirements and purpose of an analysed written composition [3, 47], including student essays. Many researchers who have analysed student essays have a) defined rhetorical categories of a student essay reflecting categories specified in the essay scoring scheme (e.g., clearly presented thesis statement, argument and counterargument) and b) provided evidential sentences that support each argument and counterargument [8, 46, 48].

Recently, researchers have begun examining rhetorical structure of entrance essays at Brazilian universities using RST [17, 42]. In this task, prospective university students were required to develop an opinion paper on an assigned topic (e.g., the effects of global warming). To observe the characteristics of an opinion essay, a seven-category rhetorical scheme was proposed in [17] including the following categories: Title, Theme, Thesis, Argumentation, Background, Conclusion, and Author. The students are required to provide a Title **s0** for their essays and introduce the essay topic in a few opening sentences that are categorised as Theme **t1**. Next, students present their stances about the essay topic, i.e., they develop Thesis **t2** sentences. The thesis is followed by the argumentation section where students discuss their positions, present and logically justify their arguments, the sentences classified as Argumentation **s2**. Finally, students provide a summary of thesis (Background, **t3**) and convey closing arguments (Conclusion, **s3**) and cite relevant Authors (**s4**). In a recent paper, Nau et al. [43] confirmed that the rhetorical scheme – which was analysed in previous work [17, 42] – is the main structure to evaluate an student essay in Brazil. However, Nau et al. [43] highlight that within all rhetorical categories, the most relevant ones are thesis (the main part of the introduction), argumentation and conclusion, as they are crucial for the cohesion of the text.

2.2 Related Works

Researchers have utilised different natural language processing and machine learning algorithms to develop approaches to analysing the essay structure at scale. The algorithms developed to date have afforded the opportunity to automatically identify a rhetorical category of an EDU in student essays of a different genre (e.g., [9, 17, 42, 46, 48]) and also to identify EDU features that are predictive of the EDU category (e.g., [48]). Following, we review these approaches.

2.2.1 Automatic detection of rhetorical categories in student essays. To date, researchers have proposed different models for automatic identification of rhetorical categories in student essays. The main methods are based on the use of natural language processing models and/or machine learning algorithms.

Several papers apply information from the Xerox Incremental Parser [49], which uses syntactic parsing, dedicated lexicons and pattern-matching rules to identify the rhetorical categories [21, 31]. These studies reached high level of accuracy, identifying the majority of the rhetorical categories in the texts. They also showed that the students were satisfied with the use of a system based on the extracted rhetorical categories as a way to provide feedback. However, this method depends entirely on natural language models, which do not have the same accuracy for different languages [4], and human-generated rules that are hard and time-consuming to be created.

Due to the lack of natural language models for different languages and the time-consuming task to manually generate rules, many studies have been using machine learning algorithm. For instance, Burstein et al. [9] labelled sentences in a dataset of argumentative and expository essays written by high-school and undergraduate students as belonging to one of the following categories: Title, Introductory Material, Thesis, Main Idea, Supporting Idea, Conclusion, and Irrelevant. The authors used this dataset to train a decision-tree model that achieved a high prediction performance, particularly in detecting sentences that provide evidence and support to claims in the essays (F1=[.89–.91]). Aiming to identify argument components and relations among them in a corpus of student essays, Nguyen and Litman [46] assigned each sentence an argument type (Major Claim, Claim and/or Premise) and connected those types via argument relations (Support or Attack). These authors created a decision model based on a topic modelling algorithm that reported a high prediction performance (F1≥.93) in identifying support relations among rhetorical components. Rakovic et al. [48] analysed sentences in a corpus of student argumentative essays to identify evidential sentences that draw upon the information borrowed from source texts. The evidential sentences were further annotated using the framework articulated from Bloom's typology [33] to distinguish between sentences that copy/paraphrase vs transform source information. The Random Forest classification algorithm developed to automatically identify evidential sentences representing the transformation of source knowledge achieved the prediction accuracy of .73. Liu et al. [36] used different machine learning algorithms (Random Forrest, SVM, Naïve Bayes and PART) to classify sentences into Reflective or Non-reflective. The main novelty of the Liu et al. [36] work is the use of the measures provided by Linguistic Inquiry and Word Count (LIWC) [54] and Academic Writing Analytics (AWA) [31] instead of using traditional content-based features. In the best-case scenario, using the Random Forest algorithm, Liu et al. [36] produced the results that reached an F1 of .799.

In the context of Brazilian college entrance essays, dos Santos et al. [17] utilised the seven-category coding framework (described in Section 2.1) to analyse the essays written in the form of opinion texts and in the Portuguese language by aspiring college students. dos Santos et al. [17] trained two different classifiers (SVM and CRF) to predict the sentence's class as Title, Theme, Thesis, Argumentation, Background, Conclusion or Author. The algorithm performed particularly well in detecting Argumentation sentences in student texts (F1=.83). Its performance in detecting other rhetorical categories that significantly contribute to an entrance essay score (i.e., Theme, Thesis, Background and Conclusion) was notably lower (F1=[.39–.66]). The study presented in [17] also highlights the advantages of using sequence-based prediction models such as Conditional Random Fields [51] to identify EDU, as the rhetorical structures in general are highly dependent on the text flow. Nau et al. [43] also proposed an analysis of rhetorical categories for Brazilian college entrance essays. In this case, the authors introduced ILP (Integer Linear Programming) to support the classification of EDUs into thesis, argumentation and conclusion. The main classifier evaluated was SVM with different kernels. The final results showed that

ILP was not relevant to improving the final result as it achieved F1 of 0.598 and the model without ILP reached 0.620.

Taken together, the aforementioned studies have demonstrated an empirical potential to automatically and accurately identify rhetorical categories in argumentative essays written by students. Not only did these models rely upon different coding frameworks for discourse analysis, but also upon different approaches to feature extraction, broadly focused on content features (e.g., lexicons of cue words [9, 46, 48]), relational features (extracted using rhetorical structure trees [9], topic modelling [46], and latent semantic analysis [48]) and ordering features (common across all the reviewed models), e.g., the relative position of the sentence within a paragraph, with some exceptions like [36]. Including content features in the feature set could, however, increase the likelihood of over-fitting ([27]), because the prediction model could become overly content-dependent when trained on the vocabulary used by a group of students. Consequently, the model may perform poorly when tested on the essay corpus developed by a different group of students, as different vocabulary may be presented in this new corpus.

To replace content features in the model, researchers studying educational discourse have recently proposed the adoption of content-independent features ([5, 18]), e.g., features generated by the Coh-Metrix ([22]) and LIWC ([54]) tools. For example, Mello et al. [42] took this line of research a step further and developed a classification model with reduced likelihood of over-fitting. Specifically, the authors utilised the same seven-category coding framework (Section 2.1) and the same data set as done in [17], and developed a group of classification models that relied upon content-independent features, i.e., the features whose values are not influenced by specific content of the text in the training corpus. Those features include dictionary-based indices of psychological processes [11], indices of text complexity and coherence [41], ordering features and features extracted from adjacent sentences. The best performing model (XGBoost) identified in the Mello et al. [42] study achieved a considerable accuracy (Cohen's $\kappa = 0.67$) in predicting rhetorical categories in Brazilian college entrance essays. Equally important, as this model completely relied upon content-independent features it holds a promise to improve generalisability of the predictive model. However, despite these initially promising results, researchers have yet to examine whether and to what extent the performance of content-independent classification models trained on one data set is maintained when the models are applied to a different data set containing essays of the same genre. This is particularly critical to ensure that the automatic classification models can perform comparably well in predicting the essay rhetorical categories across different writing styles of students and across different topics assigned every year to students taking the opinion essay entrance exams.

Another relevant lack in the literature is related to the machine learning models used. The majority of the papers presented in this section applied traditional machine learning models, which showed that SVM and Random Forest reached better results when compared to other models [17, 36, 43]. Recently, the CRF algorithm has also started to be applied [17]. However, state-of-the-art decision tree algorithms, such as AdaBoost and XGBoost, which have reached better results in the literature compared to SVM and Random Forrest [13, 14], have not been fully explored to identify rhetorical categories. Finally, we have not found any prior work that performed a generalisability analysis of the proposed methods. This evaluation is critical as, in general, the models developed are applied to different contexts. For instance, in the case of Brazilian college entrance essays, every year, a new topic is selected as the theme for the production of the essay. Without a generalisable model, it is not possible to comparatively analyse the texts written in across different academic years.

2.2.2 Linguistic features predicting rhetorical categories. Upon identifying the highest-performing and sufficiently generalizable classification model, researchers can more closely analyse linguistic features in the model to identify those features predictive of particular rhetorical categories that determine the essay score. By doing so, researchers can gain insight into characteristics of the essay that can lead to low score, a valuable information that can support instructional interventions helping prospective students improve their writing skills.

In several previous studies, researchers have examined linguistic features in the context of argumentative writing. For instance, Taguchi et al. [53] studied linguistic features in students' placement argumentative essays to distinguish between low- and high-achieving essays. The authors found that low-achieving writers tend to produce overly complex sentences, e.g., as measured by their excessive use of subordinating conjunctions (e.g., "if"). Further, McNamara et al. [40] identified syntactic complexity, lexical diversity and word frequency as the most predictive indices of the argumentative essay quality, while Yang and Sun [59] reported that the correct use of cohesive devices, e.g., for addition (e.g., "moreover"), causation (e.g., "therefore"), and temporality (e.g., "then") in writing was positively related to essay quality. Graduate students participating in Yang et al. [58] study were assigned two argumentative essays on different topics. The authors found that mean sentence length and clausal complexity predicted writing scores across topics. Similarly, Yang et al. [58]showed that students who developed longer units of text at clausal and sentential level (e.g., use of compound sentences) achieved higher scores for their writing. Finally, Liu et al. [36] presented the top-10 features used to identify rhetorical categories, which includes the use of quantifiers, adjectives, word counts, truthful and space words, among others.

2.3 Research Questions

The first goal of our research was to improve further automatic identification of rhetorical categories in Brazilian college entrance essays that researchers have previously achieved. Building upon Mello et al. [42] work, in the present study, we examined the performance and generalisability of content-independent classification models, aiming to address the existing gaps in the literature. In particular, we trained and experimentally tested the performance of the AdaBoost and XGBoost classifiers in relation to the algorithms previously used in the literature (SVM, Random Forest and CRF). To this end, we adopted the same coding framework and utilised the same data set (i.e., annotated sentences) as in [43] because the coding framework proposed is theoretically grounded [55]. This coding framework also best resembles the main structure of the Brazilian entrance essays. We thus trained each ML algorithm by using either standard content-dependent features or sequential content-independent features. Then, we examined the performance of all the algorithms to unveil whether content-independent algorithms can perform sufficiently well compared to content-dependent ones. Thus, the first research question in the current study was:

> **Research Question 1 (RQ1):**
> *What is the performance of content-independent classification models compared to sequential content-independent classification models in the classification of EDUs into the rhetorical categories in student essays?*

Next, to examine the generalisability of the sequential content-independent classification approach proposed, we tested the performance of our models on two data sets (1) the dataset created in [17] and (2) the dataset created in [43]. In particular, we performed the cross-course experiment, i.e., all the models were trained on one data set and tested on another data set, and vise versa. For completeness, we also included the content-dependent models in the analysis. As such, the second research question was formulated as:

> **Research Question 2 (RQ2):**
> *To what extent is the classification performance achieved by content-independent classification models consistent across different data sets?*

Finally, even if previous studies demonstrated the potential of using linguistic features to predict essay quality, it has not been widely adopted to identify rhetorical categories. Moreover, to our knowledge, researchers have not analysed linguistic features in the context of student essay performance in Brazilian college entrance exams. To address this gap in the literature, we examined the most important features predicting rhetorical categories in the observed essays. To this end, we utilised the best performing models developed in this study to address the final third research question:

> **Research Question 3 (RQ3):**
> *Which classification features are the most predictive of rhetorical categories of the essay EDU and how many of these features are based on the sequential approach?*

3 METHOD

3.1 Datasets

This work assessed the proposed methods using two datasets created by [17] and [43], which are composed of sentences extracted from essays written in Brazilian Portuguese containing different rhetorical categories annotated.

The dataset created by dos Santos et al. [17] (in this paper referred to as the Santos dataset) comprised 271 texts divided into 2,562 sentences. The texts were written by candidates for a university entrance exam from 2014 to 2016. Three human annotators, two with a background in computer science and one in linguistics, were responsible for the annotation process of each of the sentences in one of the following seven categories: Title, Theme, Thesis, Argumentation, Background, Conclusion, and Author. In the case of sentences reflecting multiple categories, the annotators selected the most prominent category. The agreement among the annotators reached the value of 0.78 for Cohen's κ. Table 1 presents the statistics of this dataset.

Similarly, Nau et al. [43] developed a dataset (in this paper referred to as the Nau dataset) containing 50 essays and 659 sentences extracted from the Brasil Escola website[3] in 2020. This website has the purpose of supporting students training for the Brazilian college entrance exams. Two experts in linguistics annotated each sentence of the essays into three categories: Thesis, Argumentation, and Proposal of Intervention. Moreover, sentences that do not belong to any of the above categories were considered as Non-Argumentative. The agreement between the experts reached 0.92 (accuracy) and 0.72 (Krippendorff's alpha coefficient). A third expert resolved the divergences identified between the first two annotators. Table 2 presents the statistics of the Nau dataset.

Table 1: Kappa scores and distribution of categories for the Santos dataset [17].

	Number of sentences	κ
Title	213	1.00
Theme	544	0.73
Thesis	259	0.70
Argumentation	958	0.77
Background	115	0.56
Conclusion	218	0.73
Author	255	1.00
Total	2,562	0.78

Table 2: Krippendorff's alpha coefficients and distribution of categories for the Nau dataset [43]

	Number of sentences	α
Thesis	62	0.87
Argumentation	222	0.91
Proposal of Intervention	100	0.95
Non-Argumentative	275	-
Total	659	0.92

As the categories in the datasets are different from each other, in this study, we used only three of them: Thesis, Argumentation and Conclusion. It is important to notice that according to the authors who developed the two datasets [17, 43], Thesis, Argumentation and Conclusion are the main categories in the Brazilian college entrance essays. Further, the Conclusion category from the Santos dataset and the Proposed Intervention category from the Nau dataset refers to the same rhetorical category [17, 43]; in this paper, we refer to this category as Conclusion for both datasets for consistency.

3.2 Feature Extraction

Traditionally, automatic identification of rhetorical structure has been performed based on content features, commonly adopted

[3] https://brasilescola.uol.com.br/

in text classification problems [8, 17, 46]. In this study, we compared the performance of predictive models using standard TF-IDF features and content-independent features based on linguistic resources instead (i.e., LIWC and Coh-Metrix), which have been widely adopted to classify different reflexive texts in educational settings [5, 40, 45, 53, 59] including rhetorical categories [36]. Moreover, we also proposed a new approach, called sequential content-independent features, using the features from adjacent sentences and not only from the analysed sentences to incorporate information about the flow of the rhetorical units relevant for this context [37, 38].

3.2.1 *TF-IDF Features.*

One of the most common approaches used in classification models is to adopt the traditional Term Frequency measure — Inverse Document Frequency (TF-IDF) to extract features from texts [39]. This method performs a transformation of a textual document (e.g., online discussion messages) to an array consisting of the term counts [39], in this case the TF-IDF values. The current study adopted the traditional TF-IDF technique [39].

3.2.2 *Content-independent features.*

We also evaluated the performance of features extracted by well-known linguistic tools (i.e., LIWC [54] and Coh-Metrix [41]). These features have been largely harnessed in other problems in educational research, e.g., analysis of online discussion posts [18, 45] and categorisation of feedback messages [12].

(1) **LIWC features**: The Linguistic Inquiry and Word Count (LIWC) is a textual analysis tool that computes the degree of use of different categories of words. The central component of the tool is the lexical dictionaries which contain several measures indicative of different psychological categories such as affective, cognitive, perceptual, and social [54]. In this work, we adopt the Portuguese version of LIWC proposed by [11], created based on the English version of LIWC [54], which had 73 word count categories used as features. It is important to mention that previous work had used LIWC to identify rhetorical categories [36].
(2) **Coh-Metrix features**: Coh-Metrix is a computational linguistics tool that provides measures to assess the cohesion, coherence, readability, and linguistic complexity of a text using different levels of analysis, such as lexical, syntactic, and discourse [41]. This tool has been widely used in previous studies to analyse the coherence and structure of essays (e.g., [1, 16, 35]). In the current study, we adopted the Portuguese version of Coh-Metrix used in [10] which extracted 98 features.
(3) **Ordering features**: The Rhetorical Structure Theory [37, 38] indicated that capturing the flow of ideas in a document is essential to identify text segments in a rhetorical structure model [56]. Seeking to reflect this idea, we incorporated two features that capture the order of sentences in an essay text: i) the position from the first to the last sentence (ascending order); and ii) the position of the sentences from last to first (descending order).
(4) **Features extracted from adjacent sentences**: Related research [17, 20] achieved better results by adopting a sequence-based classification approach, incorporating the context of the sentences during the classification process. Thus, the features vector of a sentence s_i incorporated its own features and those of the sentences s_{i-1} and s_{i+1} when these existed.

Considering LIWC, Coh-Metrix, and the ordering, the initial features space used in this work had a total of 173 features. As the features of the previous and subsequent sentences were also incorporated (if they existed), the final features vector of a sentence had 519 features – i.e., 3 x 173 features.

3.3 Model Selection and Evaluation

We trained several machine learning classifiers to address our research questions, traditional algorithms (Random Forest and Gaussian kernel SVM), state-of-the-art decision tree approaches (AdaBoost and XGBoost), and sequence-based classifiers (CRF). The traditional classifiers were selected based on their performance in previous studies [17, 36, 43]. Random Forest is a bagging technique that combines different decision trees generated via data sub-sampling in the training set [24]. SVM is an interactive algorithmic approach that utilises statistical methods to create a hyperplane that divides the training dataset into two categories [50]. In the case of a multi-classification problem, the outcome is a combination of several SVM classifiers. More recently, AdaBoost and XGBoost are decision tree algorithms that have demonstrated better results when compared to Random Forest [13, 14]. These algorithms use the boosting technique, an iterative process where each classification tree deals with cases that have been incorrectly classified [14, 25]. Finally, we also evaluated the performance of the Conditional Random Fields (CRF) algorithm. The literature shows the effectiveness of adopting CRF for sequence labelling problems such as the analysis of rhetorical structures and identification of named entities in textual documents [51]. In short, CRF creates a graph model to analyse the neighbourhood of the instances in the categorisation process. Thus, the CRF model considers the context of the instance to predict its category.

We used the same evaluation process performed in the previous work [17] to compare the classification results. To measure the performance of supervised machine learning algorithms, we adopted Cohen's κ [15], a metric commonly used in educational data mining and learning analytics [32, 45], and precision, recall, and F1, which are widely used metrics in the field of machine learning [2].

To address research question RQ1, we applied a 10-fold stratified cross-validation sampling to both datasets combined. It is important to highlight that we analysed sentences as EDU; however, we used cross-validation at an essay level to avoid introducing pieces of the same essay in two different sets (i.e., training and testing). In this part of the study, we assessed the performance of different predictive models using the standard TF-IDF features (content-based), content-independent features (e.g., Coh-Metrix and LIWC), and the combination of content-independent features with the features extracted from adjacent sentences.

We performed cross-training and testing on the dataset level to answer research question RQ2. Specifically, we performed the training process on one dataset and the testing on the other one and then repeated the process by changing the roles of training and testing sets. We evaluated the same algorithms and feature sets of RQ1.

Finally, to address research question RQ3, we assessed the importance of the content-independent features in terms of their contribution to the prediction of the rhetorical categories considered in the curent study. We selected the two best performing models (XGBoost and CRF) identified in the analyses performed to address the first two research questions. The Mean Decrease Gini impurity index (MDG) [7] and the Transition Feature Coefficients (TFC) [51] were adopted to estimate the importance of the individual features for XGBoost and CRF, respectively.

4 RESULTS

4.1 RQ1: Performance of content-independent features

Table 3 presents the results of the machine learning algorithms that were trained using TF-IDF, content-independent features, and content-independent features from the current and adjacent sentencesThe results revealed that the models based on TF-IDF features outperformed the models based on content-independent features when applied with SVM, Random Forrest, and AdaBoost algorithms. However, the models in best performing classifiers, i.e., XGBoost and CRF, were based on the content-independent features from current and adjacent sentences. These models, therefore, achieved the 17% higher performance than the models based on content-dependent features, measured by Cohen's κ. Finally, the combination of XGBoost and content-independent features achieved the best classification results according to all the metrics computed in the analysis. We also note that, in the case of CRF classifier, the results based on content-independent features could not be obtained, because CRF is, by design, a sequence-based model, i.e., it takes the order of sentences into account by default.

4.2 RQ2: Evaluation across different data sets

Tables 4 and 5 present the results of the cross-dataset analyses. In both cases, we repeated the same methodology of RQ1. As well, we used the same machine learning algorithms as those in the analyses for RQ1. XGBoost with content-independent features, including the features of the adjacent sentences, achieved the best classification performance: 0.69 (precision), 0.73 (recall), 0.71 (F_1) and 0.45 (Cohen's κ), when the training set was the Santos dataset and the testing set was the Nau dataset (Table 4). Similar to the results obtained for RQ1, CRF performed similarly to XGBoost. Importantly, table 4 shows that the models based on TF-IDF features did not provide much generalisability, as the values of Cohen's κ in all of these models were close to 0.

Table 5 shows the results using the Nau and Santos datasets as training and testing sets, respectively. Again, XGBoost demonstrated the best performance (precision of 0.67, recall of 0.74, F_1 score of 0.70 and Cohen's κ of 0.37) when trained with the content-independent features in combination with features of adjacent sentences. Confirming the results of the previous analysis, CRF achieved similarly strong generalisation results, showing Cohen's κ of 0.36 when trained on the combination of the content-independent features and the features of adjacent sentences. On the other hand, AdaBoost reached results closer to those of XGBoost, Cohen's κ of 0.32, when trained on the combination of the content-independent features and the features of adjacent sentences. The models based on the TF-IDF features maintained the poor performance in this analysis.

In the cross-dataset analyses, XGBoost reached values of Cohen's κ at a 'fair to moderate agreement' level [34] when trained on content-independent features and their combination with features of adjacent sentences. Also, in both cases, XGBoost and CRF reached better F1 results when compared to the other classifiers. Outcomes of the models based on TF-IDF in each of the cross-dataset analyses were considered at a 'no agreement' level [34]. This, in turn, indicated that the use of TF-IDF is not a promising option in developing generalisable classification models of rhetorical categories in Brazilian college entrance essays .

4.3 RQ3: Feature importance

To answer research question RQ3, we analysed the most important features in XGBoost and CRF classifiers, as these classifiers outperformed all the other models in our analysis. Moreover, these two classification models differ from each other in terms of their algorithmic nature, i.e, XGBoost is an instance-based classifier, whereas CRF is a sequence-based classifier.

Table 6 shows the top 20 most important features for the XGBoost classifier, as ranked based on the MDG impurity index. Here, we highlight the four key findings: (i) 14 out of the 20 features were based on the values of the previous (-1) or following (+1) sentences, and only six were related to the analysed sentence. This result shows the efficacy of the proposed approach using the sequences of features for the XGBoost classifier; (ii) the most predictive features were those related to the order of the sentences in the text (descending and ascending), which was expected as the rhetorical categories in the Brazilian college admissions essays suppose to appear in a specific order; (iii) the majority of the relevant features (11) were extracted from LIWC; and (iv) there were several features that could be related to the students' opinion or specific content related to the topic of the text in the essay (i.e., liwc.feel, liwc.anger and liwc.sexual), which is expected for this type of text.

Table 7 shows the results of the similar feature importance analysis for the CRF classifier. In this case, the features were ranked according to the Transition Feature Coefficient (TFC) measure. Similar to the results shown in Table 6, a majority of the features (11) were related to the previous (-1) or following (+1) sentences rather than the features of the current sentence. Moreover, LIWC also contributed 11 features to the list of top-20 most most important features and the sentence order was relevant (ascending). Although we can see several the similarities in the results shown across the two tables, only the same five features occurred in both Tables 6 and 7.

5 DISCUSSION

The results obtained for the automatic classification of the rhetorical categories in Brazilian Portuguese essays indicated that the proposed approach, based on sequential content-independent features, reached best performance when applied in combination with XGBoost and CRF algorithms. The values of Cohen's κ were 0.67 (XGBoost) and 0.63 (CRF) represent a moderate level agreement rate [34]. Such a result aligns with the literature as XGBoost consistently reaches better outcomes when compared to traditional machine

Table 3: Results for the analysed algorithms in terms of precision, recall, F1, and Cohen's κ.

	TF-IDF				Content-independent				Sequential content-independent			
Algorithm	**Precision**	**Recall**	**F1**	κ	**Precision**	**Recall**	**F1**	κ	**Precision**	**Recall**	**F1**	κ
SVM	0.79	0.83	0.82	0.54	0.51	0.42	0.65	0.00	0.65	0.73	0.71	0.29
Random forest	0.80	0.83	0.83	0.57	0.75	0.81	0.78	0.49	0.76	0.80	0.79	0.52
AdaBoost	0.76	0.76	0.77	0.48	0.68	0.72	0.67	0.42	0.72	0.75	0.71	0.48
XGBoost	0.80	0.80	0.81	0.56	0.82	0.82	0.82	0.65	0.83	0.84	0.84	0.67
CRF	0.79	0.78	0.79	0.53	-	-	-	-	0.81	0.81	0.81	0.63

Table 4: Results for the analysed algorithms trained using the Santos dataset and tested using the Nau dataset.

	TF-IDF				Content-independent				Sequential content-independent			
Algorithm	**Precision**	**Recall**	**F1**	κ	**Precision**	**Recall**	**F1**	κ	**Precision**	**Recall**	**F1**	κ
SVM	0.41	0.32	0.38	0.01	0.43	0.34	0.40	0.00	0.43	0.50	0.48	0.01
Random forest	0.45	0.50	0.48	0.06	0.45	0.42	0.44	0.02	0.57	0.68	0.63	0.23
AdaBoost	0.44	0.48	0.46	0.02	0.50	0.63	0.57	0.11	0.57	0.62	0.59	0.26
XGBoost	0.45	0.46	0.46	0.06	0.62	0.67	0.64	0.31	0.69	0.73	0.71	0.45
CRF	0.40	0.43	0.42	0.00	-	-	-	-	0.61	0.68	0.64	0.29

Table 5: Results for the analysed algorithms trained using Nau dataset and testing on Santos dataset.

	TF-IDF				Content-independent				Sequential content-independent			
Algorithm	**Precision**	**Recall**	**F1**	κ	**Precision**	**Recall**	**F1**	κ	**Precision**	**Recall**	**F1**	κ
SVM	0.38	0.65	0.48	0.05	0.53	0.45	0.67	0.00	0.53	0.45	0.67	0.00
Random forest	0.56	0.55	0.56	0.04	0.56	0.57	0.57	0.05	0.57	0.61	0.60	0.08
AdaBoost	0.56	0.58	0.57	0.08	0.64	0.65	0.65	0.31	0.63	0.66	0.65	0.32
XGBoost	0.56	0.55	0.56	0.04	0.65	0.74	0.69	0.32	0.67	0.74	0.70	0.37
CRF	0.37	0.54	0.48	0.02	-	-	-	-	0.68	0.74	0.71	0.36

Table 6: Top-20 most important features and their values that differentiated between the rhetorical categories with the XGboost classifier

#	Variable	Description	MDG	Thesis	Argumentation	Conclusion
1	descending	Sentence order in the essay - descending	0.035	7.31 (2.70)	5.21 (2.27)	2.00 (0.81)
2	ascending	Sentence order in the essay - ascending	0.021	3.22 (1.91)	6.37 (2.65)	9.03 (2.72)
3	liwc.space	Number of words related to physical space	0.016	0.02 (0.17)	0.03 (0.18)	0.04 (0.22)
4	-1.liwc.prep	Number of prepositions	0.014	0.53 (0.75)	0.76 (0.88)	0.87 (1.02)
5	-1.cm.DESPC	Number of paragraphs	0.013	9.44 (7.63)	11.07 (6.69)	12.10 (7.26)
6	-1.cm.DRPVAL	Syntactic pattern density: agentless passive voice density, incidence	0.012	20.07 (14.54)	23.38 (12.57)	25.75 (13.90)
7	-1.liwc.we	Number of first person plural pronouns	0.012	0.07 (0.24)	0.07 (0.24)	0.11 (0.29)
8	+1.liwc.feel	Number of words related to the perceptual processes of feeling	0.011	4.26 (2.72)	3.77 (2.30)	2.08 (2.46)
9	+1.liwc.anger	Number of words related to the affective processes of anger	0.010	0.03 (0.17)	0.03 (0.17)	0.02 (0.15)
10	+1.liwc.negemo	Number of words related to negative emotions	0.009	3.56 (2.30)	3.18 (2.00)	1.70 (2.03)
11	liwc.negemo	Number of words related to negative emotions	0.009	3.50 (1.92)	3.37 (2.13)	3.10 (1.80)
12	+1.cm.DESPL	Avg. length of paragraphs within the text	0.009	0.21 (0.55)	0.16 (0.49)	0.05 (0.24)
13	+1.liwc.relativ	Number of words related to relativity	0.009	0.06 (0.30)	0.03 (0.21)	0.01 (0.10)
14	-1.liwc.shehe	Number of third person singular pronouns	0.009	0.56 (0.78)	0.82 (0.94)	0.96 (1.06)
15	+1.cm.DESSC	Number of sentences	0.008	4.55 (3.51)	3.98 (2.82)	2.29 (2.98)
16	liwc.discrep	Number of words related to the cognitive processes of discrepancy	0.008	0.25 (0.44)	0.15(0.36)	0.13 (0.36)
17	-1.cm.SMCAUSv	Situational model: incidence score of causal verbs	0.007	0.62 (0.91)	0.59 (0.90)	0.73 (1.02)
18	-1.cm.LSASSp	LSA similarity between adjacent sentences	0.007	1.42 (1.41)	1.76 (1.57)	1.91 (1.68)
19	liwc.sexual	Number of words related to the biological processes of sexuality	0.007	0.03 (0.20)	0.01 (0.13)	0.00 (0.09)
20	-1.cm.LDTTRa	Lexical diversity, all words	0.007	10.42 (7.19)	9.51 (5.86)	5.44 (6.76)

learning algorithms [14], even for the problem of categorisation of rhetorical structure [42]. Similarly, there is the evidence in the previous work that CRF outperforms traditional machine learning algorithms for this task [17, 42].

However, the main finding related to the first research question is the effectiveness of applying content-independent features to better identify rhetorical categories (using XGBoost). To the best of our knowledge, no similar analysis has been done in the previous research. The literature reports several studies based on content-dependent features [9, 17, 43, 46, 49] and a few studies based on content-independent features [36, 42], however, a direct comparison between the two has not been performed in previous studies.

Regarding the second research question, our results (Tables 4 and 5) reveal that, as expected, the classification results of the models based on content-dependent features did not sufficiently generalize over the essays written on different topics [19], as indicated by

Table 7: Top-20 most important features and their values that differentiated between the rhetorical categories with the CRF classifier

#	Variable	Description	TFC	Thesis	Argumentation	Conclusion
1	liwc.sexual	Number of words related to the biological processes of sexuality	1.42	0.03 (0.20)	0.01 (0.13)	0.00 (0.09)
2	+1.liwc.relativ	Number of words related to relativity	1.24	0.06 (0.30)	0.03 (0.21)	0.01 (0.10)
3	cm.LSASSp	LSA similarity between adjacent sentences	0.96	0.01 (0.18)	0.02 (0.21)	0.03 (0.24)
4	-1.cm.CRFCWO1d	Content word overlap between adjacent sentences	0.91	0.00 (0.07)	0.03 (0.20)	0.03 (0.17)
5	liwc.relativ	Number of words related to relativity	0.90	0.01 (0.12)	0.05 (0.26)	0.02 (0.16)
6	+1.liwc.certain	Number of words related to the cognitive processes of certainty	0.82	0.01 (0.10)	0.02 (0.10)	0.16 (0.10)
7	+1.cm.SYNMEDwrd	Minimal Edit Distance considering all words	0.73	0.13 (0.37)	0.11 (0.34)	0.10 (0.32)
8	liwc.motion	Number of words related to motion actions	0.71	0.02 (0.17)	0.04 (0.22)	0.01 (0.10)
9	cm.CNCConfor	Number of conformative connectives	0.62	0.30 (0.52)	0.25 (0.51)	0.30 (0.63)
10	+1.cm.LSASSp	LSA similarity between adjacent sentences	0.61	0.01 (0.16)	0.02 (0.19)	0.03 (0.20)
11	+1.liwc.sexual	Number of words related to the biological processes of sexuality	0.60	0.01 (0.12)	0.02 (0.15)	0.00 (0.07)
12	cm.SYNLE	Avg. of left embeddedness words before main verb	0.58	0.23 (0.44)	0.09 (0.32)	0.10 (0.33)
13	+1.liwc.we	Number of first person plural pronouns	0.57	0.11 (0.28)	0.07 (0.23)	0.06 (0.23)
14	+1.liwc.future	Number of words related to future tense	0.53	0.14 (0.37)	0.16 (0.41)	0.03 (0.20)
15	-1.liwc.space	Number of words related to physical space	0.50	0.02 (0.16)	0.02 (0.16)	0.05 (0.23)
16	-1.liwc.number	Number of numbers	0.49	0.09 (0.09)	0.01 (0.04)	0.03 (0.05)
17	ascending	Sentence order in the essay - ascending	0.46	3.22 (1.91)	6.37 (2.65)	9.03 (2.72)
18	liwc.inhib	Number of words related to the cognitive processes of inhibition	0.45	0.03 (0.05)	0.06 (0.08)	0.00 (0.00)
19	cm.LDTTRc	Lexical diversity, all words	0.42	0.09 (0.33)	0.16 (0.41)	0.11 (0.34)
20	+1.cm.DESPL	Avg. length of paragraphs within the text	0.41	0.21 (0.55)	0.16 (0.49)	0.05 (0.24)

Cohen's κ of less than 0.1 than each of these models achieved. On the other hand, the use of content-independent features the classification performance of the models we analysed, resonating with prior research that showed benefits of using content-independent features in different educational applications [44]. Equally important, our results indicated that integration of sequential features improved performance of all the machine learning models evaluated in this study.

To answer the third research question, we investigated the feature importance in the models developed in our study. To that end, we replicated methodology proposed in previous research Barbosa et al. [5], Neto et al. [45]. Specifically, we computed the feature importance measures to identify the top 20 features for each classifier. Our results (4.3), can motivate two main conclusions: (i) the most important features, in general, were related to features extracted from previous (-1) or following (+1) sentences, which, once again, highlighted the importance of the sequential features. To the best of our knowledge, no previous work had used this approach (sequential content-independent features) to identify rhetorical categories. However, previous studies demonstrated that the idea of using sequential features and analysis are relevant for this problem [17, 30]; (ii) many of the important features reflected psychological processes, e.g., LIWC (i.e., liwc.sexual, liwc.space, liwc.feel and liwc.anger). In particular, many of these features were related to personal opinions and practical arguments [54], important rhetorical elements of a quality dissertative-argumentative essay [23]; (iii) the feature analysis performed on the results also included several Coh-Metrix features among the most relevant features (the top-20 features included 30% of Coh-Metrix measures). It is aligned with the literature that expressed the importance of Coh-Metrix to develop essay evaluation systems [16, 40, 41]..

We note the three major practical implications of our study. First, the proposed classifier, developed using sequential content-independent features and XGBoost, has been shown to automatically and with considerable accuracy identify rhetorical categories in entrance essays, promising to reduce the time that human assessors need to review and score each essay manually, because the presence of rhetorical categories, as demonstrated in previous studies [52], can predict the essay score. Furthermore, the instructions for the Brazilian college entrance essays explicitly says that 20% of the grade is related to the rhetorical structure[4]. Second, the identification of the main features for each rhetorical category could generate valuable information in creating formative feedback to guide essay revisions [26]. This is the main benefit of using white-box machine learning algorithms. This study evaluated only the top 20 features. Third, the combination of the model to categorise rhetorical structure and the identification of the most relevant features for this goal could provide the foundation for the development of learning analytic tools for supporting instructors and students in evaluating and writing better-structured essays [28, 29]. Finally, it is important to highlight that the classifier developed in this study will be used as a basis for an essay analysis system funded by the Brazilian Ministry of Education to support students' training and automatic scoring.

6 LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

We acknowledge the following limitations of the study. First, the data used in the evaluation comprised a relatively small number of texts (321 essays combining both datasets) produced by Brazilian students and with three rhetorical categories. Although this can hamper the generalizability of the method and results, several previous works evaluated with fewer data, and our findings are aligned with the literature. In future works, we intend to increase the sample size and use essays from more universities, written over different admission cycles to expand upon our results. Second, the data came from one language, Portuguese, with a relatively good number of consolidated linguistic resources and tools. We intend, as future work, to apply machine translation algorithms to identify the rhetorical structure for text written in languages with fewer resources (e.g. Arabic). Finally, this study did not intend to evaluate

[4] https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_do_enem_2020_-_cartilha_do_participante.pdf

the application of the developed model in practice to assess the satisfaction of instructors and students with a learning analytics tool based on the rhetorical structure theory. However, the development of such a tool is a promising line of future research.

ACKNOWLEDGMENTS

This research was partially supported by the fellowship 310060/2018-4 of the National Council for Scientific and Technological Development – CNPq.

REFERENCES

[1] Katherine A Abba, R Malatesha Joshi, and Xuejun Ryan Ji. 2019. Analyzing writing performance of L1, L2, and Generation 1.5 community college students through Coh-Metrix. *Written Language & Literacy* 22, 1 (2019), 67–94.

[2] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In *Mining text data*. Springer, 163–222.

[3] David Antons, Amol M Joshi, and Torsten Oliver Salge. 2019. Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management* 45, 7 (2019), 3035–3076.

[4] Arthur Barbosa, Máverick Ferreira, Rafael Ferreira Mello, Rafael Dueire Lins, and Dragan Gasevic. 2021. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 77–87.

[5] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 605–614.

[6] John D Bransford, J Richard Barclay, and Jeffery J Franks. 1972. Sentence memory: A constructive versus interpretive approach. *Cognitive psychology* 3, 2 (1972), 193–209.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] J. Burstein, D. Marcu, and K. Knight. 2003. Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18, 1 (2003), 32–39. https://doi.org/10.1109/MIS.2003.1179191

[9] Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18, 1 (2003), 32–39.

[10] Raissa Camelo, Samuel Justino, and Rafael Ferreira Leite de Mello. 2020. 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*. SBC, 179–186.

[11] Flavio Carvalho, Rafael Guimarães Rodrigues, Gabriel Santos, Pedro Cruz, Lilian Ferrari, and Gustavo Paiva Guedes. 2019. 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*. SBC, 24–34.

[12] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. How good is my feedback? a content analysis of written feedback. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 428–437.

[13] Jonathan Cheung-Wai Chan and Desiré Paelinckx. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment* 112, 6 (2008), 2999–3011.

[14] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[16] Scott A Crossley and Danielle S McNamara. 2011. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning* 21, 2-3 (2011), 170–191.

[17] Karina Soares dos Santos, Mariana Soder, Bruna Stefany Batista Marques, and Valéria Delisandra Feltrim. 2018. Analyzing the Rhetorical Structure of Opinion Articles in the Context of a Brazilian College Entrance Examination. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 3–12.

[18] Máverick Ferreira, Vitor Rolim, Rafael Ferreira Mello, Rafael Dueire Lins, Guanliang Chen, and Dragan Gašević. 2020. Towards Automatic Content Analysis of Social Presence in Transcripts of Online Discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/3375462.3375495

[19] Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 6 (2019), e1332.

[20] James Fiacco, Elena Cotos, and Carolyn Rose. 2019. Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. 310–319.

[21] Andrew Gibson, Adam Aitken, Ágnes Sándor, Simon Buckingham Shum, Cherie Tsingos-Lucas, and Simon Knight. 2017. Reflective writing analytics for actionable feedback. In *Proceedings of the seventh international learning analytics & knowledge conference*. 153–162.

[22] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher* 40, 5 (2011), 223–234.

[23] Aluizio Haendchen Filho, Hércules A do Prado, Edilson Ferneda, and Jonathan Nau. 2018. An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science* 126 (2018), 788–797.

[24] Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. Comparing automated text classification methods. *Int. J. Res. Marketing* 36, 1 (Mar. 2019), 20–38. doi: 10.1016/j.ijresmar.2018.09.009.

[25] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. 2009. Multi-class adaboost. *Statistics and its Interface* 2, 3 (2009), 349–360.

[26] Shiyan Jiang, Kexin Yang, Chandrakumari Suvarna, Pooja Casula, Mingtong Zhang, and Carolyn Rose. 2019. Applying Rhetorical Structure Theory to student essays for providing automated writing feedback. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*. 163–168.

[27] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. 2014. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*. IEEE, 372–378.

[28] Dora Kiesel, Patrick Riehmann, Henning Wachsmuth, Benno Stein, and Bernd Froehlich. 2020. Visual Analysis of Argumentation in Essays. *IEEE Transactions on Visualization and Computer Graphics* (2020).

[29] Simon Knight, Sophie Abel, Antonette Shibani, Yoong Kuan Goh, Rianne Conijn, Andrew Gibson, Sowmya Vajjala, Elena Cotos, Ágnes Sándor, and Simon Buckingham Shum. 2020. Are you being rhetorical? A description of rhetorical move annotation tools and open corpus of sample machine-annotated rhetorical moves. *Journal of Learning Analytics* 7, 3 (2020), 138–154.

[30] Simon Knight, Roberto Martinez-Maldonado, Andrew Gibson, and Simon Buckingham Shum. 2017. Towards mining sequences and dispersion of rhetorical moves in student written texts. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 228–232.

[31] Simon Knight, Simon Buckingham Shum, Philippa Ryan, Ágnes Sándor, and Xiaolong Wang. 2018. Designing academic writing analytics for civil law student self-assessment. *International Journal of Artificial Intelligence in Education* 28, 1 (2018), 1–28.

[32] Vitomir Kovanovic, Srecko Joksimovic, Dragan Gasevic, and Marek Hatala. 2014. 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.

[33] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.

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

[35] Syed Latifi and Mark Gierl. 2020. Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing* (2020), 0265532220929918.

[36] Ming Liu, Simon Buckingham Shum, Efi Mantzourani, and Cherie Lucas. 2019. Evaluating machine learning approaches to classify pharmacy students' reflective statements. In *International Conference on Artificial Intelligence in Education*. Springer, 220–230.

[37] William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

[38] William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8, 3 (1988), 243–281.

[39] Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

[40] Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication* 27, 1 (2010), 57–86.

[41] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

[42] Rafael Ferreira Mello, Giuseppe Fiorentino, Péricles Miranda, Hilário Oliveira, Mladen Raković, and Dragan Gašević. 2021. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. In *International*

Conference on Artificial Intelligence in Education. Springer, 162–167.

[43] Jonathan Nau, Rudimar Luís Scaranto Dazzi, Aluizio Haendchen Filho, and Anita Fernandes. 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*. SBC, 48–59.

[44] Valter Neto, Vitor Rolim, Anderson Pinheiro Cavalcanti, Rafael Dueire Lins, Dragan Gasevic, and Rafael Ferreira Mello. 2021. Automatic Content Analysis of Online Discussions for Cognitive Presence: A Study of the Generalizability across Educational Contexts. *IEEE Transactions on Learning Technologies* (2021).

[45] Valter Neto, Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, Rafael Dueire Lins, and Rodrigo Lins. 2018. Automated analysis of cognitive presence in online discussions written in portuguese. In *European Conference on Technology Enhanced Learning*. Springer, Springer International Publishing, 245–261. https://doi.org/10.1007/978-3-319-98572-5_19

[46] Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1127–1137.

[47] Ch Perelman. 2012. *The new rhetoric and the humanities: Essays on rhetoric and its applications*. Vol. 140. Springer Science & Business Media.

[48] Mladen Rakovic, Philip Winne, Zahia Marzouk, and Daniel Chang. [n.d.]. Automatic Identification of Knowledge Transforming Content in Argument Essays Developed from Multiple Sources. *Journal of Computer Assisted Learning* ([n. d.]).

[49] Duygu Simsek, Simon Buckingham Shum, Anna De Liddo, Rebecca Ferguson, and Ágnes Sándor. 2014. Visual analytics of academic writing. In *Proceedings of the fourth international conference on learning analytics and knowledge*. 265–266.

[50] Ingo Steinwart and Andreas Christmann. 2008. *Support vector machines*. Springer Science & Business Media.

[51] Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning* 2 (2006), 93–128.

[52] Maite Taboada and William C Mann. 2006. Applications of rhetorical structure theory. *Discourse studies* 8, 4 (2006), 567–588.

[53] Naoko Taguchi, William Crawford, and Danielle Zawodny Wetzel. 2013. What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *Tesol Quarterly* 47, 2 (2013), 420–430.

[54] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[55] Teun A Van Dijk. 1980. An interdisciplinary study of global structures in discourse, interaction, and cognition. *MacrostructuresErlbaum, Hillsdale, NJ* (1980).

[56] Teun A Van Dijk. 2019. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Routledge.

[57] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5021–5031. https://doi.org/10.18653/v1/2020.acl-main.451

[58] Weiwei Yang, Xiaofei Lu, and Sara Cushing Weigle. 2015. Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing* 28 (2015), 53–67.

[59] Wenxing Yang and Ying Sun. 2012. The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and education* 23, 1 (2012), 31–48.