# Automatic Detection of Narrative Rhetorical Categories and Elements on Middle School Written Essays

Rafael Ferreira Mello[1,3], Luiz Rodrigues[2(✉)], Erverson Sousa[1], Hyan Batista[3], Mateus Lins[3], Andre Nascimento[3], and Dragan Gasevic[4]

[1] CESAR School, Centro de Estudos e Sistemas Avançados do Recife, Recife, Brazil
rafael.mello@ufrpe.br
[2] Federal University of Alagoas, Alagoas, Brazil
luiz.rodrigues@nees.ufal.br
[3] Federal Rural University of Pernambuco, Recife, Brazil
[4] Monash University, Clayton, Australia

**Abstract.** Narrative essays enable students to express their thoughts and feelings, gain diverse perspectives, understand themselves and their world more deeply, and enhance their literary skills and cultural awareness. As such, it became a relevant textual production in educational settings. However, assessing these texts is challenging and time-intensive. Therefore, this study focuses on developing an automatic detection system for narrative structures in essays, using a range of natural language processing algorithms, including machine learning, deep learning, and large language models. The study found that BERT was more effective than other models for this task, highlighting GPT's potential in extracting narrative structures. The findings of this study have implications for educational practices, particularly in the assessment and improvement of narrative writing skills among middle school students.

**Keywords:** Natural Language Processing · Narrative Essays · Rhetorical Categories · Large Language Model

## 1 Introduction

Essays are concise literary productions embody an author's viewpoint on a specific subject [46]. When composing essays, students can establish their understanding of the subject matter and their analysis, synthesis, and organizational proficiency capabilities. In this context, narrative essays focus on the capability of students to write texts that recount personal experiences and stories, which might be based on reality or fiction [38]. Narrative productions are composed of a complex textual representation with several elements defined as narrative macrostructures: Character, Setting, Initiating Event, Plan, Action, Consequence, and Elaborated Noun Phrase [16]. Narrative essays are a key production developed for elementary and middle school students, both for classroom instruction and


A. M. Olney et al. (Eds.): AIED 2024, LNAI 14829, pp. 295–308, 2024.
https://doi.org/10.1007/978-3-031-64302-6_21

assessments [8]. They help develop basic writing skills and the capability of creative expression and storytelling proficiency [34].

However, assessing narrative essays is a challenging task, as they are comprised of several elements connected to each other. Analyzing and detecting them is time-consuming, especially in a large corpus [34,38]. From a teaching-learning perspective, this challenge is a substantial issue because feedback is prominent for effective learning experiences [42]. Therefore, instructors need tools to help them provide rapid feedback regarding narrative essays.

Previous research has shown that natural language processing (NLP) is a useful approach to overcome this issue [14]. Specifically, in assessing narrative essays, researchers started exploring ways to automate this process using NLP and machine learning (e.g., [23,44]). The detection of narrative elements from written essays has shown to be an efficient tool to assist instructors in providing useful feedback to their students [4]. In that regard, previous work has contributed to detecting narrative elements from medical and work-related essays (e.g., [27,34,39]).

Nevertheless, the corpora considered in previous studies on automated assessment of narrative essays are mostly written in English [2]. Although recent studies have started to address similar concerns based on Brazilian Portuguese essays, they evaluate different literary production [15,26] or they target specific elements [5]. Furthermore, despite recent advancements in Large Language Models (LLMs) – an advanced form of NLP optimized for text generation – have revealed their potential for numerous NLP tasks [32,43], past research has not systematically experimented with them in this context. Therefore, to the best of our knowledge, there is a lack of research detecting the broad range of narrative elements for essays in nun-English Languages based on NLP techniques.

This article addresses the gaps listed above with an empirical study on the automatic detection of narrative elements from narrative essays written in Brazilian Portuguese. For that, this paper reports on the findings of a study that explored a broad range of NLP algorithms, including classical machine learning, deep learning, and GPT-4, a state-of-the-art LLM [29]. Although much has been discussed regarding LLMs capabilities [32,43], the combination of BERT and Random Forrest classifier reached the best results in general, suggesting task-specific models' potential compared to generic models. In practice, this study could provide relevant information for teachers' feedback in the context of essay production and inform researchers on the limitations of LLMs for particular tasks compared to specific models.

The remainder of this paper is organized as follows. Section 2 discusses background and related works. In Sect. 3 we explain our methodology to address the problem raised. In Sections, we present how the work is evaluated and the models adopted. Section 4 we show the results. Finally, Sect. 5 and 6 contextualize the results in terms of practical adoption, explain the limitations of the study, and discuss future directions.

## 2 Background

### 2.1 Narrative Elements

A narrative text is a personal production that explores a series of connected events or experiences, often centered around characters and a temporal progression of actions [1]. Narrative essays are tools for organizing and transmitting knowledge, memories, and cultural values, allowing individuals to express their internal experiences and interpret the world around them. They are composed not only of events but also of intentions, conflicts, and resolutions that together form a cohesive and meaningful story. Therefore, a written narrative is not just a sequence of events but a carefully woven construction that reflects the complexity of human experience and the capacity for storytelling [18,24,40].

Specifically for education, narrative essays usually target the creation of a fictional universe or represent aspects of reality, involving elements such as plot, characters, theme, and style. In Brazil's primary education, the textual production of narratives is an essential pedagogical practice, as it develops linguistic and creative skills and fosters critical thinking and empathy in students [10]. Through narrative writing, students learn to articulate their ideas and emotions, explore different perspectives, and better understand themselves and their world. Literary narrative, therefore, is a powerful tool in education, as it not only improves students' writing and reading skills but also enriches their educational experience, promoting a deeper understanding of literature and culture [11,30].

More specifically, narrative essays comprise narrative components and narrative (or story) elements. According to [34], *components* are the main elements that compose the narrative are:

- *Events* might be formally defined as significant occurrences with consequences. Those can be real-world incidents, such as bomb explosions or the birth of an heir.
- *Participants*, also known as actors, are the *who* of the story. They are relevant to the *what* and the *why*, and are often defined as entities in NLP.
- *Time* concerns the story's temporal references, which are prominent for understanding the timeline of the narrative by anchoring each event or scene at a point in time.
- *Space* concerns geographical information within the narrative, such as references to places or locations.

In exploring the narrative's components, various *elements* can be examined. As suggested in [16], seven narrative elements are present in narrative essays, with degrees ranging from 0 (absent) to 3 (elaborated knowledge).

- *Character*: An agent performing actions. Levels: 0 (no main character or ambiguous pronouns), 1 (at least one main character with non-specific labels), 2 (a main character with a proper name), 3 (multiple main characters with proper names).

- *Setting*: Provides information about the story's location or time. Levels: 0 (no reference), 1 (a general place or time), 2 (references to a specific place or time), 3 (places using proper names or specifying a particular time).
- *Inciting Event(s)*: Motivates characters to take action. Levels: 0 (descriptions with no indication), 1 (one event without motivating action), 2 (one motivating event), 3 (two or more events motivating separate actions).
- *Internal Response*: Feelings expressed by characters about Inciting Events. Levels: 0 (no stated feelings), 1 (feelings not clearly related), 2 (feelings explicitly related), 3 (multiple instances of feelings clearly related).
- *Plan*: Thoughts stated by characters related to a decision to take action. Levels: 0 (no statement), 1 (statements about plans not necessarily related), 2 (one statement about a plan related), 3 (multiple statements about plans related).
- *Attempt*: Actions taken by characters motivated by the Inciting Event. Levels: 0 (no actions), 1 (action verbs in descriptive sentences without a clear link), 2 (action verbs in sentences clearly linked), 3 (introducing complicating actions that impede responses).
- *Consequence*: The end result of characters' actions in relation to the Inciting Event. Levels: 0 (no clear resolution), 1 (outcomes linked to other actions), 2 (one outcome related), 3 (two or more outcomes directly related).

The narrative elements considered for our study will be based on the narrative components outlined by [34] and [16]. These sources provide a comprehensive framework that will guide our analysis.

### 2.2 Related Works

This section presents prior research on the analysis of textual elements in narrative essays, aligning with the objectives of this paper. The automatic analysis of narrative essays has found extensive applications beyond educational purposes, extending its utility to other domains. For instance, Mulyana et al. [27] proposed a pipeline for extracting patient symptoms based on medical narrative texts for diagnosing mental illness. With the help of four databases of diseases and symptoms, the study focused on general text processing steps, e.g., sentence identification, scanning, and parsing; keyword extraction; and pattern matching on top of the narratives. However, the Mulyana et al. paper did not report an evaluation of the effectiveness of the pipeline for extracting symptoms from narrative texts.

Another application is identifying events from movie descriptions [39]. Based on using NLP to extract information, each event is structured with the following attributes: "who", "did what", "to whom", "where" and "when". Following those events, the authors use a Recurrent Neural Network (RNN) and the Word2vec algorithm to extract the descriptions' events. The authors reported an accuracy near 75% for real cases analysis.

Regarding the analysis of narrative essays in education, prior research has mostly explored machine learning for the automatic assessment of narrative

essays based on the English language. For instance, Somasundaran et al. [38] establish the foundation for automated evaluations of narrative quality in student essays. By exploring algorithms such as Regression Linear, Support Vector Regression based on RBF kernel, Random Forest, and Elastic Net, they automatically assessed narrative quality, achieving moderate results regarding Quadratic Weighted Kappa (QWK). In this study, the Linear Regression reached the best values on average, obtaining a QWK of 0.7.

In [19], authors similarly employed machine learning methods to predict scores for narrative elements compared to scores obtained by human assessors. Exploring features such as those extracted with *Coh-Metrix*, *TF-IDF*, GloVe embeddings, and Bidirectional Encoder Representations from Transformer (BERT), their predictive models achieved QWK scores aligned with those obtained by human assessors. Specifically, their findings showed that BERT performed up to two times better than the other methods, demonstrating the potential for automating the scoring of narrative elements.

While most studies on automated analysis of written narratives are related to the English language, recent research within the Brazilian Portuguese context also explores this theme. In [3], the authors aimed to classify Brazilian Portuguese texts in terms of the following types: narrative, essay, injunctive, and descriptive. They used classical machine learning algorithms based on linguistic features extracted with Coh-Metrix. Nevertheless, this study did not specifically analyze the feasibility of identifying specific narrative elements, such as plot and characters. The proposed approach achieved an f-score of 91.2% in the best case.

On a similar line, the study reported in [5] was focused on the issue of automatic correction of narrative texts in Brazilian Portuguese, emphasizing identifying the climax. Using an English annotated dataset translated into Brazilian Portuguese, the study investigated the use of machine learning algorithms to detect climaxes in textual productions using three traditional classification algorithms: Support Vector Machine, Random Forest, and Stochastic Gradient Descent. The article highlights Random Forest as the best-performing algorithm and suggests combining Coh-Metrix and LIWC attributes to produce the best results. The study concludes that it is possible to develop an automatic system for detecting climaxes in narratives, emphasizing the importance of carefully selecting features to improve the effectiveness of climax classification in narrative texts.

Furthermore, [37] explores the automatic scoring of student narrative essays in Portuguese, focusing on the *formal register*, which assesses aspects related to the formal grammar of Brazilian Portuguese. Different machine learning algorithms (i.e., Decision Tree, Random Forest, SVM, Extra-Tree, AdaBoost, and XGBoost) were evaluated using diverse linguistic features. The results revealed that the Extra-Tree Ensemble algorithm exhibited the best performance across all measures, with a weighted average precision of 0.557, recall of 0.566, F1-Score of 0.546, and Kappa of 0.367.

Moreover, it is important to note that prior research, such as [15,26], has also addressed the challenge of identifying rhetorical categories in Portuguese

texts through machine learning techniques. However, this research primarily concentrated on analyzing dissertative productions, a more common focus in the literature.

Lastly, although related work has extensively explored machine learning, one of the most advanced techniques at the time of writing, LLMs, has yet to be explored. LLMs are advanced NLP systems that excel in understanding and generating human-like text [43]. These models, such as GPT-3 and GPT-4, possess the remarkable capability to adapt to new tasks without explicit programming, making them versatile in handling diverse linguistic challenges [29]. LLMs find applications in various domains, from content creation and translation to question answering and code generation [32]. Their ability to comprehend context allows them to potentially revolutionize the field of education, particularly by helping in automatically identifying and analyzing story elements in narrative essays. However, as this section demonstrates, past studies have not assessed these capabilities of LLMs.

### 2.3 Research Question

To our knowledge, no prior research is specifically dedicated to evaluating the narrative rhetorical categories and elements in Brazilian Portuguese. Previous studies have primarily concentrated on analyzing grammatical errors and the automatic scoring of essays. This study tries to fill this literature gap by specifically targeting the automatic detection of structural elements in narrative essays written in Brazilian Portuguese. As such, our research question is:

**Research Question 1 (RQ1):**
*To what extent can natural language processing models accurately identify the rhetorical categories and elements of narrative essays written in Brazilian Portuguese?*

## 3 Method

### 3.1 Dataset

In this study, we used the dataset originally created for a general essay scoring problem. It encompasses 356 essays, divided into 3,262 sentences, written by mid-school students from Brazilian public schools in 2023. It is important to highlight that we did not remove students' grammatical errors and misspellings in the original text. Following previous work [15,26], we analyzed the elements of the text in the sentence level.

The initial categories chosen for our annotation process were influenced by those suggested in [16]. However, considering the relatively basic complexity of the essays, as they were written by students in the middle school, we opted to modify the categories slightly as follows:

– We kept the **Character**, **Initiating Event**, and **Consequence** as previously described;

- The category settings were divided into **Location** and **Time**;
- The categories Internal Response, Plan, and Attempt were converted into a single category named **Complication**.
- We incorporated the category **Narrator** due to the relevance of Brazilian texts described in previous work [11,30].

Two experienced educators individually coded each sentence in the essay corpus using the specified categories. The inter-rater reliability, measured by Cohen's $\kappa$, achieved a value of 0.65, indicating a substantial agreement between the coders. A third, more experienced teacher resolved the discrepancies. The 'Location' category was included only 15 times. Thus, it was excluded from this study. Table 1 details the count of positive instances for each category.

**Table 1.** Instance per category or element.

| | Number of sentences | % of the dataset |
|---|---|---|
| Narrator | 230 | 7.05 |
| Character | 300 | 9.19 |
| Time | 340 | 10.42 |
| Initiating Event | 1182 | 36.23 |
| Complication | 1852 | 56.77 |
| Consequence | 958 | 29.36 |

### 3.2 Feature Extraction

Traditionally, rhetorical structure identification has relied on content features [15,26,28,35]. Therefore, we evaluated the efficacy of predictive models using standard TF-IDF and word embedding features. TF-IDF, a fundamental technique in NLP studies, quantifies the importance of a word within a text by comparing its frequency in a specific document to its distribution across an entire corpus [25]. This method effectively highlights words that are uniquely significant in a given context. For word embeddings, we employed Global Vectors for Word Representation (GloVe) and Bidirectional Encoder Representation from Transformer (BERT). GloVe transforms each word into a 300-dimensional vector, capturing semantic relationships between words [31]. BERT, on the other hand, is a neural network designed for deep language understanding [12]. It pre-trains bidirectional representations from unlabeled text by considering the context from both sides of a word in all layers. This model, known for its robustness, continues to be a leading tool in many NLP applications.

### 3.3 Model Selection and Evaluation

To evaluate the proposed method, we considered different models available for NLP applications. Initially, we adopted a traditional Random Forest (RF) model due to its robustness for text classification problems [17]. RF operates as a bagging technique, combining multiple decision trees constructed through subsampling of data in the training set [6]. We used the RF classifier with both TF-IDF and BERT vector representations.

Secondly, we employed the BiLSTM (Bidirectional Long Short-Term Memory) architecture for this task, following its successful application in previous studies on English texts. LSTM networks are designed to learn from historical data, featuring an architecture that enables retaining previous information for use with current data. In a standard LSTM, data flows in a single direction and is processed once. Conversely, a BiLSTM model processes the data twice, analyzing it in both forward and reverse directions, enhancing its ability to capture context [36]. The architecture implemented in our experiments comprised several layers: an input layer with 30 neurons, an embedding layer with 50 neurons, a bidirectional layer with 32 cells, a dropout layer at 20% to prevent overfitting, a dense layer with 64 neurons and ReLU activation, another dropout layer at 20%, and finally, an output layer with one neuron using sigmoid activation. We trained this model over 50 epochs.

Finally, we incorporated a GPT model to explore the capabilities of a large-scale language model in this context. GPT models have gained significant recognition in academic studies for their capability to address many challenges [20,45]. Previous research shows the potential of using GPT to analyze the Essay even with a zero-shot approach [7]. In our study, we utilized the OpenAI API to access the GPT-4-turbo model, the latest version available at the time[1]

A critical factor in leveraging large language models is crafting well-structured prompts. The design of these prompts significantly impacts the model's ability to produce precise results [41]. In our study, the prompt formulation was tailored to suit the task's context, incorporating simple and direct instructions and specifying the desired output format. Table 2 displays the final prompt used for our task, translated into English[2].

**Table 2.** Prompt for GPT 4-Turbo.

| Element | Text |
|---|---|
| Context | Narrative text from mid-school students. |
| Instruction | Determine if the following text has the XXX element. |
| Output format | indicate 'yes' or 'no' in the response |

[1] OPENAI API: https://platform.openai.com/docs/api-reference/authentication.
[2] It is important to note that the original prompt was written in Portuguese.

We replicated the evaluation methodology used in prior research [15,26,35] to facilitate a comparative analysis of the model outcomes. For assessing the performance of the supervised machine learning algorithms, we employed Cohen's $\kappa$ [9], a metric widely recognized in the fields of educational data mining and learning analytics [13,33]. To ensure a robust evaluation, we conducted a 5-fold stratified cross-validation. This involved dividing the dataset into five equal parts, each mirroring the proportional representation of each class found in the original dataset.

## 4 Results

This section details the results of experiments designed to assess the effectiveness of various machine learning architectures in identifying the following narrative structure categories: *Narrator*, *Character*, *Time*, *Initiating Event* (IE), *Complication*, and *Consequence*. As mentioned before, our evaluation encompassed five distinct model configurations: TF-IDF combined with RF, GloVe integrated with BiLSTM, BERT paired with RF, BERT utilized alongside BiLSTM, and the application of GPT-4 Turbo. Each configuration was tested to measure its capability to accurately classify these narrative elements individually, providing a comprehensive outline of their respective performances in this task.

Table 3 shows the best results achieved by each algorithm using 5-fold cross-validation (as described in Sect. 3.3). The data indicates that combining BERT with RF generally yielded superior outcomes across all categories, except for one scenario (outcome) where TF-IDF + RF demonstrated better performance. For instance, in the case of the *Initiating Event* category, the BERT + RF configuration surpassed the GPT-4 model's performance by a significant margin, achieving a 32.96% higher score in terms of Cohen's $\kappa$ coefficient. This comparison highlights the distinct strengths of different algorithmic approaches in handling specific narrative structure elements.

It is important to emphasize that GPT-4 emerged as the second-best performing model in our experiments. This outcome is particularly noteworthy, considering the limited extent of prompt engineering involved. The promising results achieved by GPT-4, with minimal customization in its prompt design, underscore its potential to identify the narrative elements.

Finally, Table 3 also displays that, in general, Cohen's $\kappa$ values achieved moderate to substantial agreement levels across most categories [22], with the exception of the *Consequence* category. This finding emphasizes the effectiveness of the proposed approach. The ability of our models to consistently reach these levels of agreement across a diverse range of narrative categories demonstrates their robustness in accurately identifying and analyzing narrative structures.

**Table 3.** Comparative analysis between the four machine learning frameworks for narrative structure detection task based on Cohen's Kappa.

| Class | TF-IDF + RF | GloVe + BiLSTM | BERT + RF | BERT + BiLSTM | GPT4-Turbo |
|---|---|---|---|---|---|
| Narrator | 0.2488 | 0.3333 | **0.6624** | 0.2462 | 0.4695 |
| Character | 0.1822 | 0.1778 | **0.6255** | 0.2690 | 0.5342 |
| Time | 0.4298 | 0.4569 | **0.7237** | 0.3441 | 0.6882 |
| IE | 0.3873 | 0.3757 | **0.7039** | 0.2975 | 0.5423 |
| Complication | 0.2147 | 0.1324 | **0.5847** | 0.1389 | 0.4978 |
| Consequence | **0.2708** | 0.1432 | 0.1775 | 0.1975 | 0.1991 |

## 5 Discussion

The results obtained for the automatic identification of the rhetorical categories and elements in narrative texts written in Brazilian Portuguese essays indicated that the proposed experimentation, with BERT as features and Random Forrest as a machine learning algorithm, reached a moderate to substantial agreement [22] for almost all indicators, except for the consequence category. While no prior research has specifically addressed narrative texts, previous studies on different types of texts in Portuguese have attained a Cohen's kappa coefficient of up to 0.67 in identifying rhetorical categories [15,26,35]. This comparison emphasizes the significance and relevance of our findings in this domain.

Additionally, our research marks the first evaluation of GPT-4, a state-of-the-art LLM, for this specific task. Although GPT-4 did not outperform the combination of BERT and Random Forrest, it achieved high Cohen's kappa values, indicative of moderate agreement [22]. This outcome highlights the potential of employing GPT-4 in this context, as with further advancements in prompt engineering, there is a substantial possibility for enhancing the model's performance in this task [20,41,45].

In summary, the practical implications for research and practice can be categorized into three distinct aspects. First, the analysis of different combinations of features and machine learning algorithms, including the GPT model, demonstrated the potential to identify rhetorical categories and elements in narrative texts written in Brazilian Portuguese automatically. Second, we introduced a dataset of narrative essays written by middle school students in Brazil, which will be accessible to the research community. This represents a significant contribution, as there are few datasets with these specific characteristics available for scholarly exploration in this field. Finally, as previous research has explored, integrating the categories and elements extracted from essays into an educational tool could aid instructors and students in assessing and composing more effectively structured narrative essays [15,21].

## 6 Limitation and Future Directions

This study's primary limitations are related to the unbalanced nature of the dataset. Specific categories, such as 'place,' are underrepresented, making it challenging to train a model that can reliably detect their presence in any written essay. Additionally, the number of samples per category is highly imbalanced, necessitating the employment of undersampling or oversampling techniques to prevent the development of a biased model. In addition, the experiments were conducted with a dataset in a single language, which can reduce the generalizability of the findings. In future works, we plan to enhance the scope of our study by augmenting the sample size and incorporating essays from other languages and contexts.

In addition to the dataset issues, the machine learning models assessed in this study considered features from a single sentence in order to classify the rhetorical categories. Previous studies have demonstrated the potential of using sequential-based features and classifiers that consider the order of the sentences in the texts. Thus, in future works, we intend to evaluate the sequential method proposed by [15,26] other classifiers, such as Conditional Random Fields (CRF).

Finally, it should be noted that the primary aim of this study was not to assess the practical application of the developed model in educational settings nor to evaluate the satisfaction of instructors and students with a tool based on the extracted information. However, it is important to emphasize that this is part of an ongoing project, which includes the practical implementation of such a tool and the subsequent evaluation of its effectiveness and impact.

**Acknowledgment.** This study was partialy supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (310888/2021-2) and Acuity Insights under the Alo Grant program. Also, we would like to thank OpenAI research Program for offering the credits required for this experiment.

## References

1. Abbott, H.P.: The Cambridge Introduction to Narrative. Cambridge University Press, Cambridge (2008)
2. Bai, X., Stede, M.: A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. Int. J. Artif. Intell. Educ. 1–39 (2022)
3. Barbosa, G.A., et al.: Aprendizagem de máquina para classificação de tipos textuais: Estudo de caso em textos escritos em português brasileiro. In: Anais do XXXIII Simpósio Brasileiro de Informática na Educação, pp. 920–931. SBC (2022)
4. Batista, H., Cavalcanti, A., Miranda, P., Nascimento, A., Mello, R.F.: Classificação multi-classe para análise de qualidade de feedback. In: Anais do XXXIII Simpósio Brasileiro de Informática na Educação, pp. 1114–1125. SBC, Porto Alegre, RS, Brasil (2022). https://doi.org/10.5753/sbie.2022.225396, https://sol.sbc.org.br/index.php/sbie/article/view/22486
5. Batista, H.H., et al.: Detecção automática de clímax em produções de textos narrativos. In: Anais do XXXIII Simpósio Brasileiro de Informática na Educação, pp. 932–943. SBC (2022)

6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Brown, T.B., et al.: Language models are few-shot learners (2020)
8. Coelho, R.: Teaching writing in Brazilian public high schools. Read. Writ. **33**(6), 1477–1529 (2020)
9. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
10. Dalla-Bona, E.M., Bufrem, L.S.: Aluno-autor: a aprendizagem da escrita literária nas séries iniciais do ensino fundamental. Educ. Rev. **29**, 179–203 (2013). https://doi.org/10.1590/S0102-46982013000100009
11. Detmering, R., Johnson, A.: "Research papers have always seemed very daunting": information literacy narratives and the student research experience. Portal: Lib. Acad. **12**, 5–22 (2012). https://doi.org/10.1353/pla.2012.0004
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Ferguson, R., Clow, D., Griffiths, D., Brasher, A.: Moving forward with learning analytics: expert views. J. Learn. Anal. **6**(3), 43–59 (2019)
14. Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C.: Text mining in education. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(6), e1332 (2019)
15. Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., Gasevic, D.: 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. pp. 404–414 (2022)
16. Gillam, S.L., Gillam, R.B., Fargo, J.D., Olszewski, A., Segura, H.: Monitoring indicators of scholarly language: a progress-monitoring instrument for measuring narrative discourse skills. Commun. Disord. Q. **38**(2), 96–106 (2017)
17. Hartmann, J., Huppertz, J., Schamp, C., Heitmann, M.: Comparing automated text classification methods. Int. J. Res. Marketing **36**(1), 20–38 (2019)
18. Jago, B.J.: Chronicling an academic depression. J. Contemp. Ethnography **31**, 729–757 (2002). https://doi.org/10.1177/089124102237823
19. Jones, S., Fox, C., Gillam, S., Gillam, R.B.: An exploration of automated narrative analysis via machine learning. PLoS ONE **14**(10), e0224634 (2019)
20. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
21. Kiesel, D., Riehmann, P., Wachsmuth, H., Stein, B., Froehlich, B.: Visual analysis of argumentation in essays. IEEE Trans. Visual. Comput. Graph. (2020)
22. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics (1977)
23. Li, Z., Liu, F., Antieau, L., Cao, Y., Yu, H.: Lancet: a high precision medication event extraction system for clinical text. J. Am. Med. Inform. Assoc. JAMIA **17**, 563–7 (2010). https://doi.org/10.1136/jamia.2010.004077
24. Luo, Y.H.: Word technology and literature narrative. J. Southwest Univ. Sci. Technol. (2010)
25. Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
26. Mello, R.F., Fiorentino, G., Miranda, P., Oliveira, H., Raković, M., Gašević, D.: Towards automatic content analysis of rhetorical structure in Brazilian college entrance essays. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 162–167. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78270-2_29

27. Mulyana, S., Hartati, S., Wardoyo, R., et al.: A processing model using natural language processing (NLP) for narrative text of medical record for producing symptoms of mental disorders. In: 2019 Fourth International Conference on Informatics and Computing (ICIC), pp. 1–6. IEEE (2019)
28. 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)
29. OpenAI: Gpt-4 technical report (2023)
30. Parry, B.: Introduction: a Narrative on Narrative. Palgrave Macmillan (2013). https://doi.org/10.1057/9781137294333_1
31. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
32. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023)
33. Rodrigues, L., et al.: Question classification with constrained resources: a study with coding exercises. In: Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N., Santos, O.C. (eds.) International Conference on Artificial Intelligence in Education, pp. 734–740. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36336-8_113
34. Santana, B., Campos, R., Amorim, E., Jorge, A., Silvano, P., Nunes, S.: A survey on narrative extraction from textual data. Artif. Intell. Rev. 1–43 (2023)
35. 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
36. Siami-Namini, S., Tavakoli, N., Namin, A.S.: The performance of LSTM and BILSTM in forecasting time series. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 3285–3292. IEEE (2019)
37. da Silva Filho, M.W., et al.: Automated formal register scoring of student narrative essays written in portuguese. In: Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil. pp. 1–11. SBC (2023)
38. Somasundaran, S., Flor, M., Chodorow, M., Molloy, H., Gyawali, B., McCulla, L.: Towards evaluating narrative quality in student writing. Trans. Assoc. Comput. Linguist. **6**, 91–106 (2018)
39. Tozzo, A., Jovanović, D., Amer, M.: Neural event extraction from movies description. In: Proceedings of the First Workshop on Storytelling. pp. 60–66. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/W18-1507, https://aclanthology.org/W18-1507
40. Venkatraman, K., Thiruvalluvan, V.: Development of narratives in Tamil-speaking preschool children: A task comparison study. Heliyon **7** (2021).https://doi.org/10.1016/j.heliyon.2021.e07641
41. White, J., et al.: A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382 (2023)
42. Wisniewski, B., Zierer, K., Hattie, J.: The power of feedback revisited: a meta-analysis of educational feedback research. Front. Psychol. **10**, 3087 (2020)
43. Yenduri, G., et al.: GPT (generative pre-trained transformer) - a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions (2023)

44. Zhang, F., Fleyeh, H., Wang, X., Lu, M.: Construction site accident analysis using text mining and natural language processing techniques. Autom. Constr. **99**, 238–248 (2019)
45. Ziyu, Z., et al.: Through the lens of core competency: Survey on evaluation of large language models. In: Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum), pp. 88–109 (2023)
46. 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