



Automated Analysis of Cognitive Presence in Online Discussions Written in Portuguese

Valter Neto¹, Vitor Rolim¹, Rafael Ferreira^{1,2}(✉), Vitomir Kovanović³,
Dragan Gašević^{2,4}, Rafael Dueire Lins¹, and Rodrigo Lins¹

¹ Departamento de Computação, Universidade Federal Rural de Pernambuco,
Recife, Brazil

{valter.neto,vitor.rolim,rafael.mello,rafael.lins,
rodrigo.linsrodrigues}@ufrpe.br

² University of Edinburgh, Old College, South Bridge, Edinburgh EH8 9YL, UK
{rafael.ferreira,dragan.gasevic}@ac.ed.uk

³ University of South Australia, 160 Currie St, Adelaide, SA 5000, Australia
Vitomir.Kovanovic@unisa.edu.au

⁴ Monash University, 19 Ancora Imparo Way, Clayton, VIC 3800, Australia
dragan.gasevic@monash.edu

Abstract. This paper presents a method for automated content analysis of students' messages in asynchronous discussions written in Portuguese. In particular, the paper looks at the problem of coding discussion transcripts for the levels of cognitive presence, a key construct in a widely used Community of Inquiry model of online learning. Although there are techniques to coding for cognitive presence in the English language, the literature is still poor in methods for others languages, such as Portuguese. The proposed method uses a set of 87 different features to create a random forest classifier to automatically extract the cognitive phases. The model developed reached Cohen's κ of .72, which represents a "substantial" agreement, and it is above the Cohen's κ threshold of .70, commonly used in the literature for determining a reliable quantitative content analysis. This paper also provides some theoretical insights into the nature of cognitive presence by looking at the classification features that were most relevant for distinguishing between the different phases of cognitive presence.

Keywords: Community of Inquiry (CoI) model · Content analytics
Online discussions · Text classification

1 Introduction

The adoption of Learning Management Systems (LMSs) has increased significantly in the last few years [30]. Such systems provide resources that can enable social interactions between students, as well as between students and their teachers. Among the resources available in LMSs, asynchronous discussion forums are

widely used for encouraging student course participation, answering questions, and sharing resources [17]. Online discussions play an important role in the educational experience of students, especially in fully online learning courses, given the absence of face to face interactions.

The Community of Inquiry (CoI) model [14] emphasizes the social nature of modern online learning and it is one of the most researched and validated pedagogical model in the domain of distance education. It defines three constructs (known as presences) that shape students online learning, with the central construct being the *cognitive presence*, which captures the development of the critical and in-depth thinking skills [14] of the students. The *Quantitative Content Analysis (QCA)* method [37, 42] is widely adopted to assess the three CoI presences, making valid and reliable inferences from the analysis of textual data [5]. The CoI model defines three QCA coding schemes, one for each presence which can be used to analyze the discussion messages of the students online at the three presence levels. Although widely adopted in the social sciences within CoI community, content analysis has been primarily used for retrospection and research after the courses are over, without much impact on the actual student learning and outcomes [41]. In this regard, automated methods for text analysis commonly used within learning analytics [13] have a potential for making an assessment of CoI presences easier and less labor intensive, with the ultimate goal of using CoI model to drive instructional interventions and affect student learning outcomes [21].

There have been promising approaches for automating the assessment of cognitive presence [8, 22, 23, 31, 44], but the focus of those studies have been exclusively on English language courses, limiting their use to English-speaking countries only. Likewise, the availability of text analytics tools to languages other than English is even more limited, causing a significant deleterious effect on the accuracy of the systems developed for those languages. The different student demographics and course context within non-English courses can have a substantial effect on the predictive power of the developed analytics. The growing need for high-quality education in developing countries, implies in the need to examine how such findings can be replicated within courses in languages other than English and how analytics findings can be used for supporting students in non-English-speaking countries.

This paper describes the results of the study which examined the use of automated text analytics methods for assessing the cognitive presence from online discussion transcripts written in Portuguese. The study was based on the previous work within English-language courses [22, 24, 44] and adopted a similar classification approach, albeit with some modifications due to the differences between English and Portuguese text analytics tools. The classification method of Kovanović et al. [24] was successfully adopted showing some evidence of the potential of employing existing text analytics to non-English courses. Moreover, despite of the fact that Portuguese analysis tools and libraries are slightly less developed, the classification accuracy of 83% and Cohen's κ of .72 obtained in the experiments performed were better than the ones reported by the previous studies [22, 24, 44] showing the role of the context on the final analytics findings. The results and their implications are further discussed in this paper.

2 Background

2.1 The Community of Inquiry (CoI) Model

The Community of Inquiry (CoI) model is a widely adopted framework that describes the different facets of students' online learning [15]. Three dimensions or presences provide an overview of online learning experience: (i) *Cognitive presence* captures the development of desirable learning outcomes such as critical thinking, problem-solving, and knowledge (co-)construction [14, 16]; (ii) *Social presence* focuses on social interactions within a group of students (i.e., cohesion, affectivity, and open communication) [36]; and (iii) Teaching presence encompasses the instructors' role before (i.e., course design) and during (i.e., facilitation and direct instruction) a course [2]. This study focuses on the cognitive presence, which captures the development of critical and deep-thinking skills [14]. The cognitive presence is operationalized through a four-phase model of practical inquiry by Lipman [29]:

- 1 *Triggering event*: A problem or dilemma is identified and conceptualized. In an educational context, discussions are usually triggered by instructors; however, they can also be initiated by any participant in the discussion.
- 2 *Exploration*: The students explore the potential solutions to a given problem, typically by information seeking and brainstorming different ideas.
- 3 *Integration*: The students synthesize new ideas and knowledge by employing social (co-)construction.
- 4 *Resolution*: Finally, students solve the original dilemma or problem triggered at the beginning of the learning cycle. Here, students evaluate the newly-created knowledge through hypothesis testing, vicarious application, or consensus building.

Despite the fact that the CoI model is well established as a very effective model for assessment of social interactions in distance learning, the coding process requires a considerable amount of manual work which leads to a problem related to the scalability of its adoption [12]. The development of the CoI survey instrument [4] was one effort to reduce the need for manual content analysis of the discussion messages. However, the CoI survey instrument relies on self-reported data which makes it not applicable for real-time monitoring and guidance of student learning. Thus, automatic methods for coding are essential to enable a broader adoption of the CoI model.

2.2 Automating Cognitive Presence Analysis

Within the published literature, there have been several studies that looked at the automation of cognitive presence content analysis. Early proposals based their approach primarily on word and phrase counts [8, 31], such as the ones provided by the General Inquirer category model [40] adopted by Mcklin [31] or fully custom dictionaries adopted by Corich et al. [8]. Using such an approach Mcklin [31], the performance figures achieved 0.69 in Holsti's Coefficient

of reliability [18] and in 0.31 Cohen’s κ . Similarly, reference Corich et al. [8] reported in 0.71 Holsti’s coefficient of reliability, albeit using a sentence-level coding and assessment rather than the more widely used message level.

Some more recent studies examined the use of other different features and classifiers. Kovanović et al. [22] examined the use of a combination of bag-of-words (n-gram) approach and Part-of-Speech (POS) N-gram features for classifying cognitive presence using the Support Vector Machines (SVMs) classifier. While the authors reported 0.41 Cohen’s κ , they also pointed out at the issue of high class imbalance (lower level exploration messages are much more common than other three types of messages), as well as overfitting the data with very high number of features (more than 20,000) on a comparatively small dataset (1,747 messages). In order to address those challenges, Kovanović et al. [24] proposed the use of features based on Coh-Metrix [32], LIWC [43], LSA similarity, named entities, and discussion context [44]. Thereby, the authors reduced the feature space from more than 20,000 features to just 205 features. In their study, Kovanović et al. [24] developed a random forest classifier [6], which also allowed for the analysis of the influence of the different features on the final classification results. For example, their findings indicated that longer and more complex messages were generally more closely related to higher levels of cognitive presence, whereas question marks and first-person singular pronouns were indicative of the lower levels of cognitive presence. This work reached the best classification values (0.63 Cohen’s κ) so far reported in the literature [24].

Since the focus of this study is on examining the use of text analytics for assessing the cognitive presence online discussion messages in Portuguese, studies that addressed the CoI model within Portuguese online courses were also examined. Although, there are some studies that looked at the CoI model within Portuguese courses [3,38], there is no publication that looked at the automation of cognitive presence assessment neither in Portuguese, nor for any language other than English, to the best of the knowledge of the authors of this paper.

Table 1. Course topics by weeks.

Week	Theme	Messages (%)
1	Uses of microscopes	511 (34.06%)
2	Cell theory	400 (26.66%)
3	Genetics	314 (20.93%)
4	DNA and cloning	275 (18.35%)
Total		1,500 (100.00%)

Table 2. Distribution of cognitive presence.

ID	Phase	Messages (%)
0	<i>Other</i>	196 (13.07%)
1	Triggering event	235 (15.67%)
2	Exploration	871 (58.07%)
3	Integration	154 (10.27%)
4	Resolution	44 (2.92%)
Total		1,500 (100.00%)

3 Method

3.1 Dataset

The dataset used in the research reported here, comes from a biology undergraduate-level course offered through a fully online instructional condition at a Brazilian public university. The dataset has 1,500 discussion messages produced by 215 students over four weeks of the course (Table 1). On average, each student produced seven messages containing 89 words on average. The purpose of the online discussions was on a theme proposed by the instructor, with participation accounting for 20% of the final course mark. However, the discussions were mostly of the type question-answer rather than online debates. The whole dataset was coded by the two coders for the four levels of cognitive presence enabling for a supervised learning approach. The inter-rater agreement was excellent (percent agreement = 91.4% and Cohen's $\kappa = 0.86$). A third coder resolved the disagreements (128 in total).

Table 2 shows the distribution of the four phases of the cognitive presence, along with the category “other” which was used for messages that did not exhibit the indicators of any cognitive presence phase. The most frequent were exploration messages, accounting for more than 58% of the data, while the least frequent were resolution messages, accounting only for 2.93% of the data. The substantial difference between the frequencies of cognitive presence phases was expected [15] and also reported in the previous studies of the CoI model [22, 24].

There are several explanations for this pattern [1]. In this particular case, the forum showed characteristics of a question-answer discussion. Thus, it does seem reasonable that students will spend more time asking questions (triggering event) and especially exploring different answers (exploration). Moreover, as discussions were designed to occur between the first and the fourth week of the course, students did not typically move onto the resolution phase that early in the course.

3.2 Feature Extraction

This work follows the same approach presented by Kovanović et al. [24], in which traditional text classification features (e.g., N-gram, POS, dependency triplets) were not adopted in order to: (i) decrease the number of features, reducing the chances for over-fitting the training data; (ii) the traditional features are very “dataset dependent”, as data itself defines the classification space; (iii) N-grams and other simple text mining features are not based on any existing theory of human cognition related to the CoI model; such features can lead to models which hard to understand their theoretical meaning.

Kovanović et al. [24] evaluated 205 features mainly based on LIWC [43] and Coh-Metrix [32]. As the resources and tools for Portuguese text analytics are limited, only 87 features were explored, but all of the best ones found in [24] were included.

LIWC Features. The LIWC (Linguistic Inquiry and Word Count) tool [43] extracts a large number of word counts which are indicative of different psychological processes (e.g., affective, cognitive, social, perceptual). As there is no implementation of LIWC for Portuguese, the features extracted were the ones that: (i) reached the best results for the state-of-art cognitive presence classifier in English [24], and (ii) can be analyzed using NLP techniques (i.e., given the dictionary-based approach of LIWC, some words can only be empirically determined as representative of the psychological processes). A total of 24 features adapted from LIWC were extracted.

Coh-Metrix Features. Coh-Metrix is a computational linguistics tool that provides different measures of text coherence ((i.e., co-reference and structural cohesion) linguistic complexity, text readability, and lexical category use [32]. Coh-Metrix has been adopted in the collaborative learning domain, for example, to predict the student performance [9] and the development of social ties [20] based on the language used in the discourse. The Portuguese version of Coh-Metrix [39] has 48 different measures (while the English version has 108). It is important to mention that the features that are missing in the Portuguese version have not achieved good results in the cognitive presence classification for English.

Discussion Context Features. In order to incorporate more context to the feature space of the current study, the features proposed by Waters et al. [44] and used by [24] were included: (i) *Number of replies*: An integer variable indicating the number of responses a given message received; (ii) *Message Depth*: An integer variable showing a position of a message within a discussion tree; (iii) *Cosine similarity to previous/next message*: The idea of these features is to obtain how much the current message builds on the previously presented information; (iv) *Start/end indicators*: It uses an indicator (0/1) showing whether a message is first/last in the discussion.

The features above are relevant to the problem under study due to the process nature of the CoI model [15], in which the students' cognitive presence is viewed as being developed over time through discourse and reflection. Moreover, due to the social-constructivist view of learning in the CoI model, the different phases of the cognitive presence tend to change over time. Thus, one expected that triggering and exploration messages would be more frequent in the early stages of the discussions, while integration and resolution messages would be more common in the later stages.

Word Embedding Similarity. Kovanović et al. [24] made a parallel about the cognitive phases and the information presented in the various stages of the learning process. In summary, the triggering phase introduces a topic, while the exploration phase introduces new ideas and answers. The integration phase keeps talking about the same ideas (by constructing the meaning from the ideas

previously introduced), and resolution concludes the discussion presenting the explicit guidelines for applying knowledge constructed [33].

Due to the reasons listed above, it is beneficial to have a feature that can identify if the context of each message changes over time in a discussion. The main difference related to the original work by Kovanović et al. [24] is that the current study adopted word embeddings to represent the word similarity instead LSA. In brief, word embeddings are neural networks algorithms to translate words into numerical vectors based on their occurrences in a text [26]. Thus, the problem of identifying the relationship between words becomes a simple measure of the cosine similarity between their vectors. In the current study, the word embeddings algorithms and trained dataset available in the spaCy tool¹ were applied.

Number of Named Entities. Previous work in the literature suggested that the number named entities (e.g., named objects such as people, organizations, and geographical locations) would be different for the different phases of cognitive [14]. Exploration messages, which are characterized by the exploration of new concepts and opinions, are expected to bear more named entities than integration and resolution messages. The spaCy library² was used to extract the number of named entities.

3.3 Data Preprocessing

The first step of the data analysis performed here divided the data into training and test datasets (75% and 25% of the whole corpus, respectively), as often done in machine learning [11]. This step was performed to avoid overestimating the model performance which can occur if the model accuracy estimated on the same data as the model parameters [11] were learned. It is important to mention that stratified samples concerning coding categories (i.e., Triggering, Exploration, Integration, Resolution, and Other) were created to preserve their distribution in both train and test subsets. The split dataset included 1,125 and 375 instances for the training and test datasets, respectively (Table 3).

After the corpus partitioning, the problem of class imbalance was addressed as shown in Table 3. The imbalance can lead to negative effects on the results of the classification analyses [35]. In this step, the approach suggested by Kovanović et al. [24] was followed, using the SMOTE algorithm [7], which creates additional synthetic data points as a linear combination of the existing data points. The SMOTE processes the data points in an n -dimensional feature space (for instance $X = f_1, f_2, f_3, \dots, f_n$) of a specific class selected for resampling as follows: (i) Find K (in our case five) nearest neighbors of X belonging to the minority class chosen; (ii) Randomly select one of the identified neighbors (called Y), (iii) Generate a new synthetic data point Z as: $X + r * Y$ where r is a random number between 0 and 1.

¹ <https://spacy.io>.

² <https://spacy.io>.

Figure 1 presents the final result of the SMOTE algorithm application in the training set. The size of the codes Other, Triggering Event, and Integration, were increased 4 to 5-fold, while the Resolution category was increased 19-fold (from 34 to 646).

Table 3. Distribution of coding categories in test and train data sets.

Phase	Dataset				
	Train		Test		Total
Other	153	(13.6%)	43	(11.47%)	196 (13.07%)
Triggering event	164	(14.58%)	71	(18.93%)	235 (15.67%)
Exploration	661	(58.76%)	210	(56%)	871 (58.07%)
Integration	113	(10.04%)	41	(10.93%)	154 (10.27%)
Resolution	34	(3.02%)	10	(2.67%)	44 (2.92%)
Total	1125	(100%)	375	(100%)	1500 (100%)

3.4 Model Selection and Evaluation

There are several machine learning techniques to build supervised models. Fernández-Delgado et al. [10] performed a sizeable comparative analysis of 179 general-purpose classification algorithms over 121 different datasets identified that random forests and Gaussian kernel SVMs were the top performing algorithm. This work adopted the random forests because it is a white-box algorithm in addition to its excellent performance. This means that it is possible to evaluate the extent to which each feature contributes to the classifier [6].

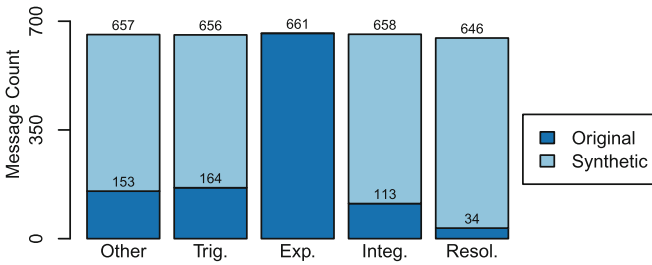


Fig. 1. SMOTE preprocessing for class balancing.

The main idea of the random forest classifier is to combine a large number of decision trees that depend on a random independently sampled vector with the same distribution for all trees. With such a mechanism, the algorithm maintains a low variance without increasing the bias [6]. It is important to mention that each tree is constructed on a different bootstrap sample of the training data,

and evaluated on the data points that were not included in the initial sample. The outcome is decided using a simple majority voting scheme.

As previously stated, the random forest algorithm allows the evaluation of the importance of the classification features. In this context, the most used measure is Mean Decrease Gini (MDG) index, which accounts for the separability of a given feature regarding the categories [6].

Finally, the two parameters used in the random forest classifiers [6] were set up: (i) `ntrree`: the number of trees generated by the algorithm; and (ii) `mtry`: the number of random features selected by each tree. Here, different values for each parameter were evaluated over the training data using 10-fold cross-validation. In both cases, the values that maximize the final performance were selected.

3.5 Implementations

The classifier was mainly coded in Python and in R programming languages. The key software packages and libraries used were:

- spaCy³, for natural language processing,
- Coh-Matrix, the Portuguese version by Scarton et al. [39],
- scikit-learn [34], for stratified sampling of test and train data,
- randomForest R package [28], for classifier development, and
- caret R package [25], for model training, selection, and validation.

4 Results

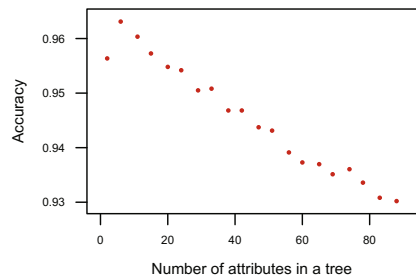
4.1 Model Training and Evaluation

Figure 2 shows the results of the tuning procedure performed in the random forest model. In the best case, the proposed classifier achieved a performance of .96 (SD = .01) classification accuracy and Cohen's κ of 0.95 (SD = .01). This result was reached with six features per decision tree on the training dataset (`mtry` = 6).

Table 4. Parameter tuning summary.

	<code>mtry</code>	Accuracy	Kappa
Min	87	0.93 (0.01)	0.91 (0.02)
Max	6	0.96 (0.01)	0.95 (0.01)
Difference		0.03	0.04

Table 5. Parameter tuning results.



³ <https://spacy.io>.

The improvement between the best- and worst-performing model was 0.03 and 0.04 for classification accuracy and κ respectively, which shows the importance of the parameter optimization in the final performance.

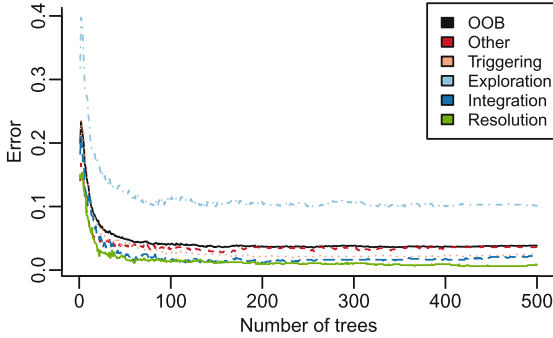


Fig. 2. Best random forest configuration performance.

Table 5 shows the performance of the random forest model using the optimal m_{try} value ($m_{try} = 6$) on the training set. There are three essential results to be analyzed in this figure: (i) the selected number of trees (500) is enough to guarantee a good classifier performance, as it stabilized with a little less than 100 decision trees; (ii) the average out-of-bag (OOB) error rate reached result

Table 6. Test data confusion matrix without the SMOTE application.

Actual	Predicted					
	Other	Triggering event	Exploration	Integration	Resolution	Error rate
Other	39	0	2	2	0	0.09
Triggering event	5	62	4	0	0	0.12
Exploration	3	2	197	8	0	0.06
Integration	1	0	24	16	0	0.60
Resolution	0	0	10	0	0	1.00

Table 7. Test data confusion matrix with the SMOTE application.

Actual	Predicted					
	Other	Triggering event	Exploration	Integration	Resolution	Error rate
Other	39	0	2	2	0	0.10
Triggering event	5	62	4	0	0	0.13
Exploration	3	2	197	8	0	0.07
Integration	1	0	24	16	1	0.61
Resolution	0	0	9	1	0	1.00

under .1, suggesting that less than 10% of the data points were misclassified; (iii) the highest error rate was observed for Exploration; this result was expected as this category was not resampled.

Tables 6 and 7 present the confusion matrix for the test data, the 25% that was left as the holdout (Table 3), before and after the application of the SMOTE algorithm. Both tables show the same result, where the error rate for the Exploration is the lowest, followed closely by the error rate for the Triggering event and Other. The tables also show that Integration and Resolution were mostly misclassified. This probably happened because these two phases had the smallest number of instances in the test dataset (Table 3), making hard for the classifier to effectively learn how to recognize messages in the two-phase.

Finally, it is important to notice that the proposed random forest model obtained .83 classification accuracy (95% CI[0.79, 0.86]) and Cohen's κ of 0.72 on the test set, which is considered a "substantial" agreement above the level of pure chance [27].

4.2 Analysis of the Feature Importance

This study also analyzed the contributions of the different features to the final performance of the classifier. Figure 3 shows the MDG scores for all classification features. It is possible to recognize that 50% of the features reached MDG score below than median (25.26) and 65% obtained an MDG score lower than the average (29.55). On the other hand, some features achieved very high MDG scores reaching 154.65 for the best feature.

Table 8 presents a detailed analysis of top twenty most relevant features. Although 87 features were evaluated, 54 had above average MDG scores; thus, due to space limitations, only the top twenty were analyzed here. It is important to note that LIWC was not used, as there is no Portuguese version of it; Thus, some features were re-implemented. The *liwc* prefix was used to refer to the features that were based on the original implementation of LIWC.

One can see that the most relevant variable was *liwc.QMark* (the number of question marks in a message), which is directly related to the Triggering phase. The average sentence length, average word per sentence, number of words and number of words bigger than six letters, number of tokens showed a similar trend, with higher values associated to Exploration and Resolution, followed closely by Integration.

Several conclusions can be drawn from the Coh-Matrix features analyzed. First, the givenness (i.e., how much information in a text is previously given) had the highest association with the higher levels of cognitive presence. The highest values for the variables of lexical diversity of the student vocabulary (VOCD and content words) were found to "other" messages. Finally, the variables related to content words and type to token ratio reached the highest values for Other and Triggering.

Regarding the features based on LIWC, they were mainly based on quantitative values (number of articles, prepositions, quantifiers, and pronouns) achieving the highest values in the exploration and resolution phases.

Finally, the variable related to the position of the message within a discussion obtained the highest values for other and triggering. This result is not usual, and the design of the discussion (debate and question-answers with a large number of the instructor’s interventions) can justify it. Most of the triggering messages were posted by the instructor trying to encourage the engagement of the students. Integration and Resolution also reached high values due to the fact that these phases usually happens after triggering and exploration messages.

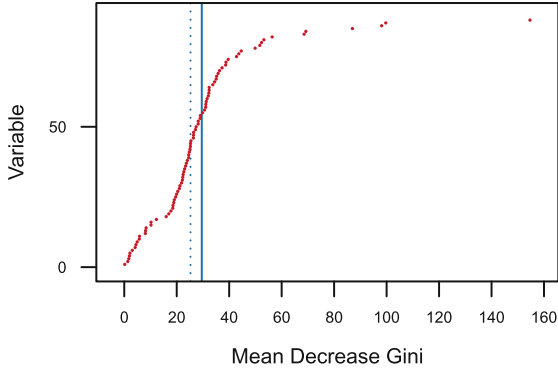


Fig. 3. Feature importance by Mean Decrease Gini (MDG) measure. Dotted blue line shows median MDG score (25.26), while solid blue line shows average MDG score (29.55).

5 Discussion

The evaluation of the automatic classification of cognitive presence over the testing dataset showed that the features based on LIWC and Coh-Metrix are effective to classify forums message in Portuguese. Cohen’s κ of 0.72 represents a “substantial” inter-rater agreement [27], and it is above the 0.70 Cohen’s which is the CoI research community commonly used as the threshold limit required before coding results are considered valid. The optimization of the *mtry* parameter (i.e., the number of attributes used in each tree of the forest) improved the final result for 0.04 Cohen’s κ and 0.003 classification accuracy (Table 4). Although the authors of this paper did not find any other related work which performed a similar analysis to compare, it is important to mention that the approach presented here reached accuracy results better than the classifiers of cognitive presence developed for English [22, 24, 44].

This study conducted a detailed analysis of the features used. First, the model was trained on only 87 features and did not use a bag-of-words vector as an attribute. Thus, the chances of over-fitting the training data decrease substantially. To draw any future conclusions about the generalizability of the classifier, it will be important to apply it to different subject domains and pedagogical

Table 8. Twenty most important features for distinguishing between cognitive presence phases and their values in different cognitive presence phases.

#	Variable	Description	MDG	Other	Cognitive presence phase			
					Triggering	Exploration	Integration	Resolution
1	liwc.QMark	Number of question marks	154.66	0.07 (0.36)	1.17 (0.80)	0.17 (0.81)	0.08 (0.37)	0.20 (0.82)
2	cm.AveSen	Average sentence length	99.69	5.98 (3.48)	8.15 (4.17)	25.3 (15.4)	23.5 (14.0)	25.7 (10.8)
3	message.depth	Position within discussion	98.12	2.65 (1.22)	2.58 (1.17)	1.52 (1.01)	2.56 (1.17)	2.18 (1.63)
4	liwc.6Word	Number of words bigger than six letters	86.96	3.04 (2.65)	5.17 (6.63)	40.4 (36.4)	17.6 (17.4)	39.9 (28.8)
5	cm.WPerSen	Avg. word per sentence	69.23	8.31 (9.53)	10.5 (6.29)	27.3 (17.9)	25.8 (16.6)	27.2 (11.0)
6	liwc.Art	Number of articles	68.58	0.86 (1.31)	1.81 (2.19)	13.6 (12.8)	5.64 (5.85)	12.7 (10.2)
7	cm.Tokens	Number of tokens	56.39	11.5 (10.98)	19.1 (20.9)	127 (111)	60.0 (51.6)	131 (95.8)
8	cm.Giveness	Average givenness of each sentence	53.24	0.47 (0.18)	0.54 (0.19)	0.73 (0.13)	0.73 (0.13)	0.69 (0.11)
9	liwc.PreP	Number of prepositions	52.33	1.16 (1.78)	1.92 (2.74)	17.1 (15.8)	7.32 (6.42)	17.9 (12.4)
10	cm.NumWord	Number of words	51.67	12.0 (12.89)	19.1 (20.5)	126 (111)	59.8 (52.2)	131 (95.3)
11	liwc.Conj	Number of conjunctions	49.89	0.35 (0.75)	0.97 (1.36)	6.80 (6.73)	3.56 (3.85)	7.43 (5.57)
12	cm.MContWord	Min. among content words frequency	44.58	520 (123)	213 (69.2)	69.3 (72.3)	39.1 (10.3)	38.4 (9.15)
13	liwc.Verb	Number of verbs	43.66	1.38 (1.69)	2.79 (3.44)	18.3 (16.4)	9.27 (7.89)	19.3 (14.6)
14	cm.LDVOCD	Lexical diversity, VOCD	42.79	0.69 (0.17)	0.59 (0.12)	0.42 (0.07)	0.46 (0.08)	0.43 (0.07)
15	cm.LDContW	Lexical diversity, content words	39.62	0.65 (0.15)	0.57 (0.11)	0.40 (0.06)	0.43 (0.05)	0.43 (0.07)
16	cm.TTR	Type to token ratio	38.79	0.95 (0.13)	0.95 (0.07)	0.74 (0.11)	0.84 (0.10)	0.76 (0.12)
17	liwc.Quant	Number of quantifiers	38.67	0.56 (0.90)	0.43 (0.81)	3.13 (3.13)	1.85 (1.95)	3.95 (3.57)
18	cm.ContWord	Content words frequency	37.29	730 (186)	629 (109)	587 (57.5)	599 (65.0)	590 (46.4)
19	liwc.3Pron	No. of pronouns in third person singular	36.33	0.04 (0.19)	0.13 (0.41)	1.55 (2.13)	0.67 (1.02)	1.68 (1.91)
20	cm.PronNP	Mean pronouns per noun phrase	35.81	0.02 (0.07)	0.04 (0.11)	0.02 (0.05)	0.05 (0.07)	0.01 (0.02)

contexts. Second, the results indicated that a small subset of features had highly predictive indicators of the different phases of cognitive presence (Fig. 3).

It is important to highlight that the most relevant classification indicators (Table 8) were aligned with the theory of cognitive presence [19]. Higher levels of cognitive presence were related to messages that are: (i) longer, with more words and sentences; (ii) complex, with complex words (words bigger than 6 letters) and longer sentences; (iii) have lower lexical diversity, as shown here by two measures of lexical diversity; (iv) have higher givenness of the information; (v) use more third-person singular pronouns; (vi) use fewer question marks. The conclusions drawn above are consistent with the findings of previous studies, for instance, 45% of the top 20 features found in the current study match those found by Kovanović et al. [24]. Future research is needed to better understand the reasons behind the differences in contributions of the features across different studies.

Finally, one can see that the Other category produces indicators with values close to the triggering phase. The Other category had messages with general requests, solicitation, or course exception rather to contribute towards knowledge construction about topics discussed. Such a category had large diversity in relation to other messages (as seen in lexical diversity and TTR features) and tended to be more informal (with fewer words, and sentences). Besides that, Other messages occurred more towards the end of a discussion, which is expected as many students would use their final post for thanking each other for their contributions.

6 Final Remarks

This paper has two main contributions. First, a new classifier to code students' transcripts on the level of cognitive presence for messages written in Portuguese was proposed. The developed approach obtained 83% accuracy and Cohen's κ of 0.72 which is considered substantial agreement above the level of pure chance [27]. This result shows the potential to provide an automated system for coding cognitive presence in Portuguese.

Second, a detailed relevance analysis of the proposed features was presented, which were mainly based on Coh-Metrix and LIWC. In such a context, the experiments performed showed that long and complex messages, along with bigger givenness and more use of third-person singular were related to higher levels of cognitive presence. Higher lexical diversity and a greater number of question marks were associated with lower levels of cognitive presence. Such conclusions corroborate the results of the related work [24].

The main limitations of the approach presented here are related to the dataset. First, the collected data was from a single study domain (i.e., biology) with discussions designed with a particular pedagogical purpose (i.e., question-answer discussion) from the same course at a Portuguese speaking university. Thus, the study may not be entirely representative of the different interactions that can lead to different cognitive presence messages. Second, the dataset size

and unbalanced categories, although consistent with the findings in the literature, may affect the performance of the classifier.

Along the lines for further work, the authors plan to test the generalization of the classifier in another education context (i.e., blended vs. fully online vs. MOOC; and undergraduate vs. graduate) and the effectiveness of the proposed features to other languages (e.g., Spanish).

References

1. Akyol, Z., Arbaugh, J.B., Cleveland-Innes, M., Garrison, D.R., Ice, P., Richardson, J.C., Swan, K.: A response to the review of the community of inquiry framework. *Int. J. E-Learn. Distance Educ.* **23**(2), 123–136 (2009)
2. Anderson, T., Rourke, L., Garrison, D.R., Archer, W.: Assessing teaching presence in a computer conferencing context. *J. Asynchronous Learn. Netw.* **5**, 1–17 (2001)
3. de Araújo, E.M., de Oliveira Neto, J.D.: Avaliação do pensamento crítico e da presença cognitiva em fórum de discussão online utilizando a análise estatística textual. In: *Proceedings of International Conference on Engineering and Computer Education*, vol. 8, pp. 113–117 (2013)
4. Arbaugh, J., Cleveland-Innes, M., Diaz, S.R., Garrison, D.R., Ice, P., Richardson, J.C., Swan, K.P.: Developing a community of inquiry instrument: testing a measure of the community of inquiry framework using a multi-institutional sample. *Internet High. Educ.* **11**(3–4), 133–136 (2008). <https://doi.org/10.1016/j.iheduc.2008.06.003>
5. Bauer, M.W.: Content analysis. An introduction to its methodology-by Klaus Krippendorff from words to numbers. Narrative, data and social science-by roberto franzosi. *Br. J. Sociol.* **58**(2), 329–331 (2007)
6. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
8. Corich, S., Hunt, K., Hunt, L.: Computerised content analysis for measuring critical thinking within discussion forums. *J. E-learn. Knowl. Soc.* **2**(1), 1–8 (2006)
9. Dowell, N.M., Skrypnik, O., Joksimovic, S., Graesser, A.C., Dawson, S., Gašević, D., Hennis, T.A., de Vries, P., Kovanovic, V.: Modeling learners' social centrality and performance through language and discourse. *International Educational Data Mining Society* (2015)
10. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res.* **15**(1), 3133–3181 (2014)
11. Friedman, J., Hastie, T., Tibshirani, R.: *The Elements of Statistical Learning*. Springer Series in Statistics, vol. 1. Springer, New York (2001). <https://doi.org/10.1007/978-0-387-21606-5>
12. Gašević, D., Adesope, O., Joksimović, S., Kovanović, V.: Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *Internet High. Educ.* **24**, 53–65 (2015). <https://doi.org/10.1016/j.iheduc.2014.09.006>
13. Gašević, D., Kovanović, V., Joksimović, S.: Piecing the learning analytics puzzle: a consolidated model of a field of research and practice. *Learn. Res. Pract.* **3**(1), 63–78 (2017). <https://doi.org/10.1080/23735082.2017.1286142>

14. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking, cognitive presence, and computer conferencing in distance education. *Am. J. Distance Educ.* **15**(1), 7–23 (2001). <https://doi.org/10.1080/08923640109527071>
15. Garrison, D.R., Anderson, T., Archer, W.: The first decade of the community of inquiry framework: a retrospective. *Internet High. Educ.* **13**(1–2), 5–9 (2010)
16. Heo, H., Lim, K.Y., Kim, Y.: Exploratory study on the patterns of online interaction and knowledge co-construction in project-based learning. *Comput. Educ.* **55**(3), 1383–1392 (2010). <https://doi.org/10.1016/j.compedu.2010.06.012>
17. Hew, K.F., Cheung, W.S.: Attracting student participation in asynchronous online discussions: a case study of peer facilitation. *Comput. Educ.* **51**(3), 1111–1124 (2008)
18. Holsti, O.R.: *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley Pub. Co., Reading (1969)
19. Joksimovic, S., Gasevic, D., Kovanovic, V., Adesope, O., Hatala, M.: Psychological characteristics in cognitive presence of communities of inquiry: a linguistic analysis of online discussions. *Internet High. Educ.* **22**, 1–10 (2014)
20. Joksimović, S., Kovanović, V., Jovanović, J., Zouaq, A., Gašević, D., Hatala, M.: What do cMOOC participants talk about in social media?: a topic analysis of discourse in a cMOOC. In: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pp. 156–165. ACM (2015)
21. Kovanović, V., Gašević, D., Hatala, M.: Learning analytics for communities of inquiry. *J. Learn. Anal.* **1**(3), 195–198 (2014)
22. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M.: Automated cognitive presence detection in online discussion transcripts. In: *Proceedings of the Workshops at the LAK 2014 Conference Co-Located with 4th International Conference on Learning Analytics and Knowledge (LAK 2014)*, Indianapolis, IN (2014). <http://ceur-ws.org/Vol-1137/>
23. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M., Siemens, G.: Content analytics: the definition, scope, and an overview of published research. In: Lang, C., Siemens, G., Wise, A., Gašević, D. (eds.) *Handbook of Learning Analytics and Educational Data Mining*, pp. 77–92. SoLAR, Edmonton (2017). <https://doi.org/10.18608/hla17.007>
24. Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., Siemens, G.: Towards automated content analysis of discussion transcripts: a cognitive presence case. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK 2016)*, pp. 15–24. ACM, New York (2016)
25. Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al.: *Caret: classification and regression training*. R package version 4 (2017)
26. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: *International Conference on Machine Learning*, pp. 957–966 (2015)
27. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977)
28. Liaw, A., Wiener, M., et al.: Classification and regression by random forest. *R News* **2**(3), 18–22 (2002)
29. Lipman, M.: *Thinking in Education*. Cambridge University Press, New York (1991)
30. McGill, T.J., Klobas, J.E.: A task technology fit view of learning management system impact. *Comput. Educ.* **52**(2), 496–508 (2009)
31. Mcklin, T.E.: *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Ph.D. thesis, Atlanta, GA, USA (2004). aAI3190967

32. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, Cambridge (2014)
33. Park, C.L.: *Replicating the use of a cognitive presence measurement tool* (2009)
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: *Scikit-learn machine learning in Python*. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
35. Rosé, C., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: *Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning*. *Int. J. Comput. Support. Collab. Learn.* **3**(3), 237–271 (2008)
36. Rourke, L., Anderson, T., Garrison, D.R., Archer, W.: *Assessing social presence in asynchronous text-based computer conferencing*. *J. Distance Educ.* **14**(2), 50–71 (1999). <http://www.ijede.ca/index.php/jde/article/view/153>
37. Rourke, L., Anderson, T., Garrison, D.R., Archer, W.: *Methodological issues in the content analysis of computer conference transcripts*. *Int. J. Artif. Intell. Educ. (IJAIED)* **12**, 8–22 (2001)
38. Rozenfeld, C.C.D.F.: *Fóruns online na formação crítico-reflexiva de professores de línguas estrangeiras: uma representação do pensamento crítico em fases na/pela linguagem*. *Alfa Rev. Linguíst. (São José do Rio Preto)* **1**, 35–62 (2014)
39. Scarton, C., Gasperin, C., Aluisio, S.: *Revisiting the readability assessment of texts in Portuguese*. In: Kuri-Morales, A., Simari, G.R. (eds.) *IBERAMIA 2010. LNCS (LNAI)*, vol. 6433, pp. 306–315. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16952-6_31
40. Stone, P.J., Dunphy, D.C., Smith, M.S.: *The general inquirer: a computer approach to content analysis* (1966)
41. Strijbos, J.W.: *Assessment of (computer-supported) collaborative learning*. *IEEE Trans. Learn. Technol.* **4**(1), 59–73 (2011)
42. Strijbos, J.W., Martens, R.L., Prins, F.J., Jochems, W.M.: *Content analysis: what are they talking about?* *Comput. Educ.* **46**(1), 29–48 (2006)
43. Tausczik, Y.R., Pennebaker, J.W.: *The psychological meaning of words: LIWC and computerized text analysis methods*. *J. Lang. Soc. Psychol.* **29**(1), 24–54 (2010)
44. Waters, Z., Kovanović, V., Kitto, K., Gašević, D.: *Structure matters: adoption of structured classification approach in the context of cognitive presence classification*. In: Zuccon, G., Geva, S., Joho, H., Scholer, F., Sun, A., Zhang, P. (eds.) *AIRS 2015. LNCS*, vol. 9460, pp. 227–238. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28940-3_18