Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case

Vitomir Kovanović
School of Informatics
The University of Edinburgh
Edinburgh, UK
v.kovanovic@ed.ac.uk

Srećko Joksimović
Moray House School of Education
The University of Edinburgh
Edinburgh, UK
s.joksimovic@ed.ac.uk

Zak Waters
Queensland University of Technology
Brisbane, Australia
z.waters@qut.edu.au

Dragan Gašević
Moray House School of Education and School of Informatics
The University of Edinburgh
Edinburgh, UK
dgasevic@acm.org

Kirsty Kitto
Queensland University of Technology
Brisbane, Australia
kirsty.kitto@qut.edu.au

Marek Hatala
School of Interactive Arts and Technology
Simon Fraser University
Burnaby, Canada
mhatala@sfu.ca

George Siemens
LINK Research Lab
University of Texas at Arlington
Arlington, USA
gsiemens@uta.edu

ABSTRACT

In this paper, we present the results of an exploratory study that examined the problem of automating content analysis of student online discussion transcripts. We looked at the problem of coding discussion transcripts for the levels of cognitive presence, one of the three main constructs in the Community of Inquiry (CoI) model of distance education. Using Coh-Metrix and LIWC features, together with a set of custom features developed to capture discussion context, we developed a random forest classification system that achieved 70.3% classification accuracy and 0.63 Cohen's kappa, which is significantly higher than values reported in the previous studies. Besides improvement in classification accuracy, the developed system is also less sensitive to overfitting as it uses only 205 classification features, which is around 100 times less features than in similar systems based on bag-of-words features. We also provide an overview of the classification features most indicative of the different phases of cognitive presence that gives an additional insights into the nature of cognitive presence learning cycle. Overall, our results show great potential of the proposed approach, with an added benefit of providing further characterization of the cognitive presence coding scheme.

Keywords

Community of Inquiry (CoI) model, content analysis, content analytics, online discussions, text classification

LAK '16, April 25 - 29, 2016, Edinburgh, United Kingdom

ISBN 978-1-4503-4190-5/16/04...$15.00

DOI: http://dx.doi.org/10.1145/2883851.2883950

1. INTRODUCTION

Online discussions are commonly used in modern higher education, both for blended and fully online learning [42]. In distance education, given the absence of face to face interactions, online discussions represent an important component of the whole educational experience. This is especially important for the social-constructivist pedagogies which emphasize the value of social construction of knowledge through interactions and discussions among a group of learners [2]. In this regard, the Community of Inquiry (CoI) model [22, 23] represents perhaps one of the best researched and validated models of online and distance education, focused on explaining important dimensions – also known as *presences* – that shape students' online learning experience.

The most commonly used approaches to the analysis of online discussion transcripts are based on the quantitative content analysis (QCA) [12, 54, 51, 15]. According to Krippendorff [37] content analysis is *"a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use"*[p18]. In the case of the study presented in this paper, contexts is online learning environments. QCA is a well defined research technique commonly used in social science research, and it makes use of specifically designed coding schemes to analyze text artifacts with respect to the defined research goals and objectives. For instance, the CoI model defines a set of coding schemes which are used by the educational researchers to assess the levels of three CoI presences.

In the domain of educational research, QCA of student discussion data have been mainly used for the retrospection and research after the courses are over without an impact on the courses' learning outcomes [53]. In the field of content analytics [36] – which focuses on building analytical models based on the learning content including student-produced content such as online discussion messages – there have been some attempts to automate some of those coding schemes. Most notable are the efforts of McKlin [44] and Corich et al. [11] on automation of the CoI coding schemes, which served

as a starting point for our research in the area [35, 62]. One of the main challenges for automation of content analysis is the fact that the most important constructs from the educational perspective (e.g., student group learning progress, motivation, engagement, social climate) are latent constructs not explicitly present in the discussion transcripts. This means the assessment of these constructs requires human interpretation and judgment.

This paper presents the results of a study that explored the use of content analytics for automating content analysis of student online discussions based on the CoI coding schemes. We focused on automation of the content analysis of *cognitive presence*, one of the main constructs in the CoI model. By building upon the existing work in the fields of text mining and text classification and our previous work in this area [35, 62], we developed a random forests classifier which makes use of a novel set of classification features and provides a classification accuracy of 70.3% and Cohen's κ of 0.63 in our cross validation testing. In this paper, we describe the developed classifier and the adopted classification features. We also report on the findings of the empirical evaluation of the classifier and critically discuss the findings.

2. BACKGROUND WORK

2.1 The Community of Inquiry (CoI) model

The Community of Inquiry (CoI) model is a widely researched model that explains different dimensions of social learning in online learning communities [22, 23]. Central to the model are the three constructs, also known as *presences*, which together provide a comprehensive understanding of learning processes [22, 23]:

1) **Cognitive presence** which is the central construct in the CoI model and describes different phases of student knowledge construction within a learning community [23].
2) **Social presence** captures different social relationships within a learning community that have a significant impact on the success and quality of the learning process [50].
3) **Teaching presence** explains the role of instructors during the course delivery as well as their role in the course design and preparation [3].

The focus of this study is on the analysis of cognitive presence, which is defined by Garrison et al. [23] as *"an extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication."*[p11]. Cognitive presence is grounded in the constructivist views of Dewey [14] and is "the element in this [CoI] model that is most basic to success in higher education" [22, p89]. Cognitive presence is operationalized by the practical inquiry model [23], which defines the following four phases:

1) **Triggering event**: In this phase, an issue, dilemma or problem is identified. In the case of a formal educational context, those are often explicitly defined by the instructors; however, they can also be initiated by the other discussion participants [23].
2) **Exploration**: This phase is characterized by the transition between the private world of reflective learning and the shared world of social construction of knowledge [23]. Questioning, brainstorming and information exchange are the main activities which characterize this phase [23].
3) **Integration**: In this phase, students move between reflection and discourse. The phase is characterized by the synthesis of the ideas generated in the exploration phase. The synthesis ultimately leads to the construction of meaning [23]. From a teaching perspective, this is the most difficult phase to detect from the discussion transcripts, as the integration of ideas is often not clearly identifiable.
4) **Resolution**: In this phase, students resolve the original problem or dilemma that started the learning cycle. In the formal educational setting, this is typically achieved through a vicarious hypothesis testing or consensus building within a learning community [23].

The CoI model defines its own multi-dimensional content analysis schemes [22, 23] and 34-item likert-scale survey instrument [5] which are used for the assessment of the three presences. The model has gained a considerable attention in the research community resulting in a fairly large number of replication studies and empirical validations (for an overview see [24]) including the studies about the interaction dynamics between the three presences [25]. In general, the model has been shown to be robust, and its coding scheme exhibits sufficient levels of inter-rater reliability for it to be considered a valid construct [24].

While the CoI model has been proven to be a very useful model for assessment of the social distance learning, there are several practical issues that still remain open. First, the use of the CoI coding schemes requires a substantial amount of manual work, which is very time consuming and requires trained coders. For example, to code the dataset used in this study, two experienced coders spent around 130 hours each to manually code 1,747 messages [26]. The coding process started with the calibration of the use of the coding scheme which was then followed by the independent coding, and finally reconciliation of the coding disagreements.

One major consequence of manual coding of messages in the CoI model is that it has been used mostly for research purposes and not for the real-time monitoring of students' learning progress and guiding instructional interventions. This is not unique to the CoI model and is very common with most of content analysis schemes used in education. The lack of automated content analysis approaches has been identified by Donnelly and Gardner [15] as one of the main reasons why transcript analysis techniques have had almost zero impact on educational practice. The development of the CoI survey instrument [5] is one attempt to eliminate, or at least to lessen the need for the manual content analysis of discussion transcripts. Still, the instrument is based on self-reported survey data, which makes it not so suitable for the real-time monitoring and guidance of student learning.

In order to enable for a broader adoption of the CoI model, the coding process needs to be automated and this is precisely the goal of the current study. While this study focuses on automation of coding online discussion transcripts for the levels of cognitive presence, a more general goal is to automate coding for all three presences, which would enable for a more comprehensive view of social learning phenomena and the development of more sophisticated social learning environments [60]. This in turn could be used by the instructors to inform their interventions leading to better achievement of learning objectives. From the standpoint of self-regulated learning research [7] – a major theory in contemporary education – in order to regulate their own learning effectively, learners need real-time feedback, which is an *"inherent catalyst"* for all self-regulated activities [7]. By providing learners with timely feedback on their own learning and the learning of their peers, they would be in a position to better regulate their own learning activities.

2.2 Automating Cognitive Presence Analysis

Several studies have investigated automating content analysis using the cognitive presence coding scheme. A study by McKlin [44] describes a system built using feed-forward, back-propagation artificial neural network that was trained on a single semester worth of discussion messages (N=1,997). The classification features were the counts of words in the one of the 182 different word categories as defined in the General Inquirer category model [52]. McKlin [44] also used a binary indicator whether a message is a reply to another

message, as triggering events are more likely to be the discussion starters and thus not replies to other messages. Finally, McKlin [44] defined custom categories of words and phrases, which are thought to be indicative of the different phases of cognitive presence and included count of words in those categories as additional classification features. For example, "indicative words" category contains "compared to", "I agree", "that reminds me of", and "thanks" as it is hypothesized that integration messages would contain larger number of these phrases in order to connect the message with the previously given information. Unfortunately, these additional coding categories are very briefly described and thus is not possible to replicate them and evaluate their usability in future studies. McKlin's findings show that classification system overgeneralized the exploration phase and under-generalized the integration phase. Furthermore, given the very low frequency of messages in the resolution phase (i.e., $< 1\%$ and only 3 messages in total in their data set), the neural network developed by McKlin simply ignored the resolution category and never predicted the resolution phase for any message in the corpus. Overall, they reported Holsti's Coefficient of Reliability [30] of 0.69 and Cohen's κ of 0.31, which show some potential of the proposed approach with much room for improvement in order to reach reliability levels commonly found among two independent coders – usually Cohen's κ of at least 0.70 [28].

Following the work of McKlin [44], a study by Corich et al. [11] presented ACAT, a very general classification framework that can support any coding scheme besides cognitive presence which is also based on word count features. In order to use ACAT, users are required to provide a set of labeled training examples, which are used for training of classification models. Furthermore, as ACAT does not specify a particular set of word categories that are used as classification features, users are required to provide definitions (i.e., category name and list of words) that are used as classification features. Interestingly, the use of the ACAT system is also evaluated on the problem of coding cognitive presence of the CoI model. However, instead of classifying each message to one of the four phases of cognitive presence, Corich et al. [11] classified each *sentence* of each message to four cognitive presence levels. This poses some theoretical challenges as the CoI coding schemes are originally designed to be used for message-level content analysis. The dataset used by Corich et al. [11] consists of 484 sentences originating from 74 discussion messages and they report Holsti's coefficient of reliability of 0.71 in their best test case. However, given that their report did not provide sufficient details about the classification scheme used in terms of the specific indicators for each category of cognitive presence, nor did it discuss the types of features that were used for classification, it is hard to evaluate the significance of their results.

Besides the studies by McKlin [44] and Corich et al. [11], we should also mention our previous work in this domain. A study by Kovanović et al. [35] investigated the use of Support Vector Machines (SVMs) [59] classification for the automation of cognitive presence coding using a bag-of-words approach based on the N-gram and Part-of-Speech (POS) N-gram features. Using a 10-fold cross-validation, a classification accuracy of 0.41 Cohen's κ was achieved – which is higher than values reported in the previous studies [44, 11].

Several challenges related to the classification of online discussion massages based on cognitive presence were observed in our existing work [35]. First, the distribution of classes in the used dataset (i.e., phases of cognitive presence) was uneven, which is in agreement with the findings commonly reported in the literature [24]. This poses some challenges to the classification accuracy. This was already seen in the McKlin [44] study whose classifier completely ignored the resolution phase (as only three messages were coded as being in resolution phase). Secondly, the use of bag-of-words features (i.e., n-grams, POS n-grams, and back-off n-grams) creates a very large feature space (i.e., more than 20,000 features) relative to the number of classification instances (i.e., 1,747) which poses challenge of over-fitting. Next, the use of bag-of-words features makes the classification system highly domain dependent, as the space of bag-of-words features is defined based on the training set. For instance, a classification system trained on a introductory programming course would likely have a bigram feature `java programming` which is highly specific to a particular domain and would impede the performance of the classifier in other domains. Finally, given that each message belongs to a discussion and represents a part of the overall conversation, the context of the previous messages in the discussion thread is very important. For example, given the structure and cyclic nature of inquiry process, it is highly unlikely that a discussion would start with a resolution message, or that the first response to a triggering message will be an integration message [26]. These "dependencies" between discussion messages are not taken into the account when each message is classified independently of other messages in the discussion.

In order to address the challenge of isolated classification of discussion messages, Waters et al. [62] developed a structured classification system using conditional random fields (CRFs) [38]. This classifier does a prediction for the whole sequence of messages within a discussion, taking into the account orderings of messages within a discussion thread. Using a 10-fold cross-validation, the developed classifier achieved Cohen's κ of 0.48 which is significantly higher than 0.41 Cohen's κ reported by [35], showing a promise of the structured classification approach. However, there are still couple of unresolved issues which warrant further investigation. First of all, although the classification accuracy is improved, it is still far below the Cohen's κ of 0.7 which is considered a norm for assessing the quality of the coding in the CoI research community [28]. Secondly, CRFs are an example of black-box classification method [27] that are hard to interpret, which limits their potential use for *understanding* how cognitive presence is captured in the discourse.

3. METHOD

3.1 Data set

The dataset used in this study is the same dataset that was used in studies by Kovanović et al. [35] and Waters et al. [62]. The data comes from a masters level, and research-intensive course in software engineering offered through a fully online instructional condition at a Canadian open public university. The dataset consists of six offerings of the course between 2008 and 2011 with the total of 81 students that produced 1,747 discussion messages (Table 1). On average, each offering of the course had $\sim$ 13-14 students ($SD = 5.1$) that produced on average $\sim$ 291 messages, albeit with a large variation in the number of messages per course offer ($SD = 192.4$). The whole dataset was coded by the two expert coders for the four levels of cognitive presence enabling for a supervised learning approach. The inter-rater agreement was excellent ($percent\ agreement = 98.1\%$, Cohen's $\kappa = 0.974$) with a total of only 33 disagreements.

Table 2 shows the distribution of four phases of cognitive presence. In addition to the four categories of cognitive presence, we included the category "other", which is used for messages that did not exhibit signs of any phase of cognitive presence. The most frequent messages were exploration messages (39% of messages), while the least frequent were the resolution messages (6% of messages). This large difference between the frequencies of the four phases was expected. It is consistent with the previous studies of cognitive presence [25], which found that a majority of students were not pro-

Table 1: Course offerings statistics

	Student count	Message count
Winter 2008	15	212
Fall 2008	22	633
Summer 2009	10	243
Fall 2009	7	63
Winter 2010	14	359
Winter 2011	13	237
Average (SD)	13.5 (5.1)	291.2 (192.4)
Total	81	1,747

Table 2: Distribution of cognitive presence phases

ID	Phase	Messages	(%)
0	Other	140	8.0%
1	Triggering Event	308	17.6%
2	Exploration	684	39.2%
3	Integration	508	29.1%
4	Resolution	107	6.1%
	Average (SD)	349.4 (245.7)	20.0% (10.0%)
	Total	1,747	100%

gressing to the later stages of integration and resolution. While there are various interpretations for this pattern, including the validity of the model, the design and expectations of the courses – i.e., not requiring students to move to those phases – seems to be the most compelling reason, as shown by its growing acceptance in the literature [24]. Psychologically, if students are going through the four phases of the practical inquiry model that underlies the cognitive presence construct, it does seem reasonable that students will spend more time exploring and hypothesizing different solutions, before they could come up with a final resolution [1, 26]. Moreover, as discussions were designed to occur between the third and the fifth week of the course, students did not typically move to the resolution phase this early in the course. Specifically, the discussions were organized to provide the students with opportunities to discuss ideas that would inform the individual research projects that they planned for the later stages of the course.

3.2 Feature Extraction

While the majority of the previous work related to text classification is based on lexical N-gram features (e.g., unigrams, bigrams, trigrams) and similar features (e.g., POS bigrams, dependency triplets), we eventually decided not to include N-gram and similar features described in the Kovanović et al. [35] study for several reasons. First of all, the use of those features inflates the feature space, generating thousands of features even for small datasets. This strongly increases the chances for over-fitting the training data. Secondly, the use of those features is also very "dataset dependent", as data itself defines the classification space. Thus, it is hard to define a fixed set of classification features in advance, as the particular choice of words in the training documents will define what features are used for classification (i.e., what N-gram variables are extracted). Finally and most importantly, given that N-grams and other simple text mining features are not based on any existing theory of human cognition related to the CoI model, it is hard to understand what they might theoretically mean. Given that our goal is also to *understand* how cognitive presence is captured within discourse, we focused our work on extracting features which are strongly theory-driven and based on empirical studies. In total, we extracted 205 classification features which are described in the reminder of this subsection.

3.2.1 LIWC features

In this study, we used the LIWC (Linguistic Inquiry and Word Count) tool [57], to extract a large number of word counts which are indicative of different psychological processes (e.g., affective, cognitive, social, perceptual). Our previous research [33] showed that different linguistic features operationalized through the LIWC word categories offer distinct proxies of cognitive presence.

In contrast to extracting N-grams, which produce a very large number of independent features, LIWC provides us with exactly 93 different word counts which are all based on extensive empirical research [58, cf.]. LIWC features essentially "merge" related – and domain-independent – N-gram features together to produce more meaningful classification features. We used the 2015 version of the LIWC software package, which also provides four high-level aggregate measures of i) analytical thinking, ii) social status, confidence, and leadership, iii) authenticity, and iv) emotional tone.

3.2.2 Coh-Metrix features

For extraction of features for classification we also used Coh-Metrix [29, 45], a computational linguistics tool that provides 108 different metrics of text coherence (i.e., co-reference, referential, causal, spatial, temporal, and structural cohesion), linguistic complexity, text readability, and lexical category use. Coh-Metrix has been extensively used a large number of studies to measure subtle differences in different forms of text and discourse and is currently used by the Common Core initiative to analyze learning texts in K-12 education [45].

Coh-Metrix has been previously used in the domain of social learning to measure the student performance [16] and development of social ties [32, 34] based on the language used in the discourse. For example, a study by Dowell et al. [16] showed that characteristics of the discourse – as measured by Coh-Metrix – were able to account for 21% of the variability in the performance of active MOOC students. Students performed significantly better when then engaged in exploratory-style discourse, with the high levels of deep cohesion and the use of simple syntactic structures and abstract language. With the goal of the existing CoI content schemes to prescribe different indicators of important socio-cognitive processes in the discourse, the use of Coh-Metrix provides a valuable set of metrics that can be easily extracted and used for automation of the CoI coding schemes.

3.2.3 Discussion context features

Drawing on the study by Waters et al. [62], we also focused on incorporating more context information in our feature space. Thus, we included all features (except unigrams) which were used in the Waters et al. study. Those included:

- *Number of replies*: An integer variable indicating the number of replies a given message received.
- *Message Depth*: An integer variable showing a position of message within a discussion.
- *Cosine similarity to previous/next message*: The rationale behind these features is to capture how much a message builds on the previously presented information.
- *Start/end indicators*: Simple 0/1 indicator variables showing whether a message is first/last in the discussion.

As the CoI model – from the perspective of educational psychology – is a process model [24], students' cognitive presence is viewed as being *developed over time* through discourse and reflection. Therefore, in order to reach higher levels of cognitive presence students need to either: i) construct knowledge in the shared-world through the exchange of a certain number of discussion messages, or ii) construct knowledge in the their own private world of reflective learning. Given the social-constructivist view of learning in the CoI

model, we can expect that the distribution of messages exhibiting the characteristics of the different phases of cognitive presence will tend to change over time, as the students progress through those phases. Thus, we can expect that triggering and exploration messages will be more frequent in the early stages of the discussions, while integration and resolution messages will be more common in the later stages.

3.2.4 LSA similarity

Messages belonging to different phases of cognitive presence are characterized with various socio-cognitive processes [23]. The triggering phase introduces a certain topic in a tentative form, presenting a concept(s) that might not be completely developed, while the exploration phase further elaborates on various approaches to the inquiry initiated in the triggering phase. More precisely, the exploration phase introduces new ideas, divergent from the community, or even several contrasting topics within the same message [49]. On the other hand, the integration phase assumes a continuous process of reflection and integration, which leads to the construction of meaning from the introduced ideas [23]. Finally, the resolution phase presents explicit guidelines for applying knowledge constructed through the inquiry process [23, 49]. Based on these insights, we assumed that information presented in the various stages of the learning process might have an important influence on message comprehension. Still, given the differences among the learners and their learning habits, we did not expect this to be manifested as a general rule, but more as a slight tendency which would be useful in combination with the other classification features.

Following the approach suggested by Foltz et al. [20], we used LSA with the sentence as a unit of analysis to define a single variable `lsa.similarity`, which represents the average sentence similarity (i.e., coherence) within a message. As LSA determines the coherence based on the semantic relatedness between terms (i.e., terms that tend to occur in a similar context) [13], we first had to define a semantic space in which the similarity estimates are given. Having in mind that different discussions might relate to the different concepts, we decided to create a separate semantic space for each discussion. We identified the most important concepts from the first message in a discussion with a semantic annotation tool TAGME [19] and then each identified concept was linked to an appropriate Wikipedia page from which we extracted information about that concept [19]. Given that previous studies [55, 21] showed that Wikipedia can be used for estimation of semantic similarity between different concepts, we used information from the extracted pages to construct the semantic space on which LSA similarity of the concepts is calculated.

3.2.5 Number of named entities

Based on the work described in [47] and our previous study [35], we hypothesized that messages belonging to the different phases of cognitive presence would contain different count of named entities (e.g., named objects such as people, organizations, and geographical locations). The basis for this is taken from the definition of the cognitive presence construct [23]. Exploration messages are characterized by the brainstorming and exploration of new ideas, and thus, those messages are expected to contain more named entities than integration and resolution messages. Given the subject of the course in which the data for this study were collected, we extracted from each message a number of entities that are related to the computer science category of Wikipedia by using the DBPedia Spotlight annotation tool [46].

3.3 Data preprocessing

As the first step in our analysis, we addressed the problem of different number of messages in five classification categories (i.e., four phases of cognitive presence and "other"). The imbalance of different classes can have very negative effects on the results of the classification analyses [56]. Generally speaking, there are two possible ways of addressing this problem [8]: i) cost-sensitive classification, in which different penalties are assigned for misclassification of instances from different categories (higher penalties for smaller classes), and thus forcing the algorithm to put more emphasis on properly recognizing smaller classes; and ii) resampling methods, either by oversampling smaller classes, undersampling large classes, or through a combination of these two approaches. Given that cost-sensitive classification is used typically for two class problems ("positive" vs. "negative"), where correctly classifying one of the classes is the primary goal of the classifier (i.e., patients with a disease, fraudulent banking transaction), it makes sense to assign different misclassification costs as correctly identifying "negative" class is not important. However, in our case, we are equally interested in all five classes (four cognitive presence categories and the other messages), as they represent different phases in student learning cycles and it is not immediately clear whether misclassification of resolution messages is "worse" than misclassification of triggering event messages. Thus, in our study, we used resampling techniques and in particular a very popular SMOTE algorithm [9], which is a hybrid approach that combines oversampling the minority class with undersampling of the majority class.

One interesting property of SMOTE is that instead of simply resampling minority class instances – which would generate simple copies of the existing data points – it generates new *synthetic* instances which are "similar" to the existing instances but not exactly the same. For example, in n-dimensional feature space, for every data point ($X = \{f_1, f_2, ...f_n\}$) of the class C_i that is selected for resampling, SMOTE:

1) Find K (in our case five) nearest neighboring instances from the class C_i. As the distances between original C_i data points are known in advance, the list of K nearest neighbors for all instances in C_i class are calculated and stored in $N \times K$ matrix (where N is the number of data points in the C_i class).
2) Randomly picks one of the identified neighbors (Y).
3) Generates a new data point Z as:

$$Z = X + rand(0, 1) * Y$$

where $rand(0, 1)$ is a function returning a random number between 0 and 1.

Figure 1 shows the results of applying SMOTE to our dataset. As our original dataset consists of 1,747 messages, the class distribution would be uniform if each of the classes contained approximately 350 messages (i.e., $1,747/5 \sim 350$). Thus, we first user SMOTE oversampling procedure explained previously to generate additional 210, 42, and 243 instances of "Other", "Triggering", and "Resolution" classes, respectively. This increased the total number of messages in each of these three classes to 350 messages in total. We then undersampled messages in "Exploration" and "Integration" categories to create a smaller groups of also 350 messages. Hence, we removed 334 "Exploration" messages and 158 "Integration" messages, to produce smaller groups of also 350 messages in total. Overall, after applying SMOTE the new dataset consists of 1,750 messages, with each of the five categories of messages represented with exactly 350 messages.

Besides compensating for class imbalance problem, we also removed the two duplicate features that were provided by both LIWC and Coh-Metrix: i) the total number of words in a message, and ii) the average number of words in a sentence. We decided to remove LIWC values and use only the ones provided by Coh-Metrix.

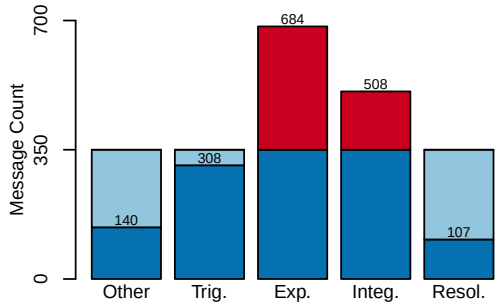

Figure 1: SMOTE preprocessing for class balancing. Dark blue – original instances which are preserved, light blue – synthetic instances, red – original instances which are removed.

The primary reason for using Coh-Metrix features is consistency, as there are some small differences in how those two systems process corner cases (e.g., hyphenated words, interpunction signs) and given that Coh-Metrix provides additional set of metrics (e.g., number of sentences, number of paragraphs) we wanted to use consistent calculations for all of the included metrics.

3.4 Model Selection and Evaluation

To build our classifier, we used random forests [6], a state-of-the art tree-based classification technique. A large comparative analysis of 179 general-purpose (i.e., not domain-specific, offline, and unstructured) classification algorithms on 121 different datasets used in the previously published studies by Fernández-Delgado et al. [18] found that random forests were the top performing classification algorithm, only matched by Gaussian kernel SVMs. Random forests are ensemble tree-based method that combines bagging (bootstrap aggregating) with the idea of random-subspace to create a robust classification system which has low variance without increasing the bias [18]. Random forests work by creating a large number of trees and then the final prediction is decided using the majority voting scheme. Each tree is constructed on a different bootstrap sample (sub-sample of the same size with repetition) and evaluated on datapoints that did not enter the bootstrap sample (in general, around one third of the training dataset size). In addition, each tree does not use the complete feature set, but has a *random* selection of N attributes (i.e., a subspace) which are then used for growing an individual tree without any pruning. Random forests are widely used technique that can handle large datasets with thousands of features.

It is important to note that random forests can also be used to measure importance of individual classification features. While importance of individual classification features might be calculated in many different ways [41], one popular measure is *Mean Decrease Gini (MDG)* which is based on the reduction in *Gini impurity* measure. Generally speaking, Gini impurity index measures how much the data points of a given tree node belong to the same class (i..e, how much they are "clean"). For every internal (split) node we can measure the decrease in Gini impurity, which shows how useful a given tree node is for separating the data (i..e, how much it reduces the impurity of the resulting groups of data). For random forests, MDG measure for a feature X_j is calculated as a mean decrease in Gini impurity of all tree nodes where a given feature X_j is used.

As there are two parameters used for configuration of random forests (i.e., `ntree` – number of trees constructed, and `mtry` – the number of randomly selected features), we used a cross-validation to select the optimal random forest parameters. As the performance of random forests typically stabilizes after a certain number of trees are built, we decided to build a large ensemble of 1,000 trees to make sure that convergence is reached. Thus, we focused on selecting optimal number of features used in every three (i.e., `mtry` parameter). We used a 10-fold cross validation and repeated it 10 times in order to reduce variability and get more accurate estimates of cross validated performance. In each run of the cross validation, we examined 20 different values for the `mtry` parameter: {2, 12, 23, 34, 44, 55, 66, 76, 87, 98, 108, 119, 130, 140, 151, 162, 172, 183, 194, 205}. The exact set of these values is obtained by using the `var_seq` function from R's `caret` package.

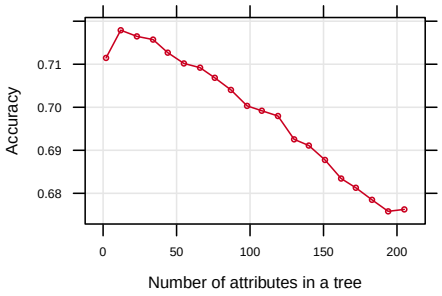

Figure 2: Random forest parameter tuning results.

Before training and evaluating our classification models, we split data to 75% for model training and 25% for testing. We used stratified sampling, so that class distribution in both sub-samples is the same. We selected the best `mtry` value using the 10 repetitions of the 10-fold cross validation and then reported the classification accuracy of the best performing model on the testing data.

3.5 Implementation

We implemented our classifier in the R and Java programming languages using several software packages:

- for feature extraction we used Coh-Metrix [45, 29] and LIWC 2015 software packages [58],
- for developing random forest classifier, we used the `randomForest` R package [40],
- for running repeated cross validation and aggregating model performance, we used the `caret` R package [31],
- for running the SMOTE algorithm we used the Weka [63] Java package, and
- for calculation of LSA similarity measure, we used the Text Mining Library for LSA (TML)[1].

The complete dataset for the study and source code of the implementation is publicly available at `github.com/kovanovic/lak16_classification` repository.

3.6 Limitations

The major limitations of our approach are related to the size of our data set. Although we have six course offerings, they are all from the same course at a single university, and together with the particular details of adopted pedagogical and instructional approach they might potentially have an effect on the generalizability of our classification model. Thus, in our future work, we plan to test the generalization power of our classifier on a different dataset, which would preferably also account for other important confounding variables recognized in research of the CoI model such as subject domain [4], level of education (i.e., undergraduate vs. graduate) [25], and mode of instruction (blended vs. fully online vs. MOOC) [61].

4. RESULTS

4.1 Model training and evaluation

Figure 2 shows the results of our model selection and evaluation procedure. The best classification accuracy of 0.72 ($SD = 0.04$) and 0.65 Cohen's κ ($SD = 0.05$) was obtained with `mtry` value of 12, which means that each decision tree takes into the account only

[1] tml-java.sourceforge.net

Table 3: Random forest parameter tuning results

	mtry	Accuracy	Kappa
Min	194	0.68 (0.04)	0.59 (0.04)
Max	12	0.72 (0.04)	0.65 (0.05)
Difference		0.04	0.06

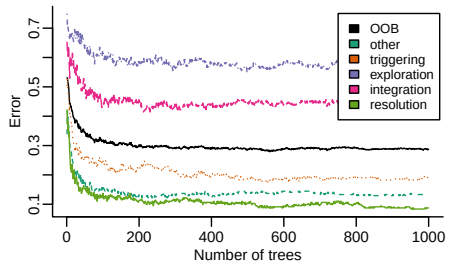

Figure 3: Best random forest configuration performance.

12 out of 205 features. The difference between the best- and worst-performing configurations was 0.06 Cohen's κ (Table 3), which suggest that parameter optimization plays an important role in the final classifier performance. Looking at the best performing configuration (Figure 3), we can see that the use of 1,000 trees in an ensemble resulted in reasonably stable error rates, with an average out-of-bag (OOB) error rate of 0.29, (i.e., an average misclassification rate for all data points in cases when they were non used in bootstrap samples). As expected, the highest error rates were associated with the undersampled classes (i.e., exploration and integration) and the smallest with the classes that were most heavily oversampled (i.e., resolution and "other").

Following the model building, we evaluated its performance on the hold-out 25% of the data. Our random forest classifier obtained 70.3% classification accuracy (95% CI[0.66, 0.75]) and 0.63 Cohen's κ which were significant improvements over 0.41 and 0.48 reported in Kovanović et al. [35] and Waters et al. [62] studies, respectively. Table 4 shows the confusion matrix obtained on the testing dataset. We can see that the most significant misclassifications are between exploration and integration messages which are hardest to distinguish. This is already witnessed in the [62] where most of the misclassifications were related to exploration and integration messages.

4.2 Variable importance analysis

Figure 4 shows the variable importance measures for all the 205 classification features. The median MDG score was 4.43, with the most of the features having smaller MDG scores, and only few features having very high MDG scores. Table5 shows the values of top 20 variables based on their MDG scores and their average values in each class (i.e., cognitive presence phase). We can see that the most important variable was the `cm.DESCWC`, i.e., the number of words in a message; that is, the longer the message was, the higher the chance of the message was to be in the later phases of the cognitive presence cycle. Also, the number of paragraphs, number of sentences, and

Table 4: Confusion matrix for the best performing model

	Predicted				
Actual	Other	Triggering	Explorat.	Integrat.	Resolut.
Other	**79**	2	2	2	2
Triggering	5	**67**	9	6	0
Exploration	9	15	**35**	27	1
Integration	2	2	23	**44**	16
Resolution	0	0	4	2	**81**

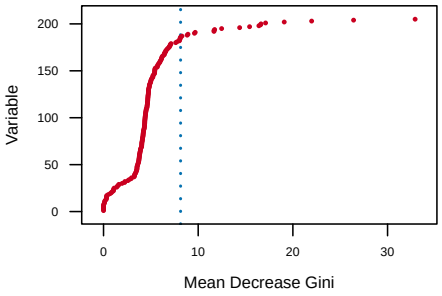

Figure 4: Variable importance by Mean Decrease Gini measure. Blue line separates top twenty features.

average sentence length showed similar trends, with higher values being associated with the later phase of cognitive presence.

The most important Coh-Metrix features were related to lexical diversity of the student vocabulary with the highest lexical diversity being displayed by "other" messages. Standard deviation of the number of syllables – which is an indicator of the use of words of different lengths – had the strongest association with the triggering event phase. In contrast, the givenness (i.e., how much of the information in text is previously given) had the highest association with the resolution phase messages. Finally, the low Flesch-Kincaid Grade level readability score and the low overlap between verbs used had the strongest association with "other" messages (i.e., messages without traces of cognitive presence).

The most important LIWC features were i) the number of question marks used, which was strongly associated with the triggering event phase, ii) the number of first person pronouns, which was highly associated with the other (i.e., non-cognitive presence) messages, and iii) use of money-related words, which is mostly associated with the integration and resolution phases.

Message context features also scored high, with message depth being higher for the later stages of cognitive presence, and highest for "other" messages. A similar trend was observed for similarity with the previous message, which was highest for the integration and resolution messages and lowest for the triggering event messages. In contrast, similarity with the next message and number of replies were highest for triggering events and lowest for the "other" messages. It is interesting to note that both LSA similarity and the number of named entities obtained high MDG scores. The number of named entities was the second most important feature and was highly associated with the later stages of the cognitive presence cycle. A similar trend was also observed for LSA similarity however, its importance was much lower.

5. DISCUSSION

Based on the testing results of the developed classifier, we can see that the use of the LIWC and Coh-Metrix features, together with a small number of thread-based context features could be used to provide reasonably high classification performance. The obtained Cohen's κ value of 0.63 falls in the range of "substantial" inter-rater agreement [39], and is just slightly below the 0.70 Cohen's κ which is the CoI research community commonly used as a threshold value for that is required before coding results are considered valid. We can also see that the parameter tuning plays an important role in optimizing the classifier performance, as the different classifier configurations obtained results different up to 0.05 Cohen's κ and 0.04% classification accuracy (Table 3).

Given that the same dataset is used as in the [35] and [62] studies, it is possible to directly compare the results of the classification algorithms. The obtained Cohen's κ is 0.15 and 0.22 higher than the ones reported by Waters et al. [62] and Kovanović et al. [35], respectively. Furthermore, the resulting feature space is much smaller,

Table 5: Twenty most important variables and their mean scores for messages in different phases of cognitive presence

#	Variable	Description	MDG*	*Cognitive presence phase*				
				Other	**Triggering**	**Exploration**	**Integration**	**Resolution**
1	cm.DESWC	Number of words	32.91	55.41 (61.06)	80.91 (41.56)	117.71 (67.23)	183.30 (102.94)	280.68 (189.62)
2	ner.entity.cnt	Number of named entities	26.41	13.44 (15.36)	21.67 (10.55)	28.84 (16.93)	44.75 (24.85)	64.18 (32.54)
3	cm.LDTTRa	Lexical diversity, all words	21.98	0.85 (0.12)	0.77 (0.09)	0.71 (0.10)	0.65 (0.09)	0.58 (0.09)
4	message.depth	Position within discussion	19.09	2.39 (1.13)	1.00 (0.90)	1.84 (0.97)	1.87 (0.94)	2.00 (0.68)
5	cm.LDTTRc	Lexical diversity, content words	17.12	0.95 (0.06)	0.90 (0.06)	0.86 (0.08)	0.82 (0.07)	0.78 (0.07)
6	cm.LSAGN	Avg. givenness of each sentence	16.63	0.10 (0.07)	0.14 (0.06)	0.18 (0.07)	0.21 (0.06)	0.24 (0.06)
7	liwc.QMark	Number of question marks	16.59	0.27 (0.85)	1.84 (1.63)	0.92 (1.26)	0.58 (0.82)	0.38 (0.55)
8	message.sim.prev	Similarity with previous message	16.41	0.20 (0.17)	0.06 (0.13)	0.22 (0.21)	0.30 (0.24)	0.39 (0.19)
9	cm.LDVOCD	Lexical diversity, VOCD	15.43	12.92 (33.93)	28.99 (50.61)	53.57 (54.68)	83.47 (43.00)	97.16 (28.95)
10	liwc.money	Number of money-related words	14.38	0.21 (0.69)	0.32 (0.74)	0.32 (0.75)	0.65 (1.12)	0.99 (1.04)
11	cm.DESPL	Avg. number of paragraphs sent.	12.47	4.26 (2.98)	6.37 (2.76)	7.49 (4.11)	10.17 (5.64)	14.05 (8.88)
12	message.sim.next	Similarity with next message	11.74	0.08 (0.14)	0.34 (0.40)	0.20 (0.22)	0.22 (0.24)	0.22 (0.23)
13	message.reply.cnt	Number of replies	11.67	0.42 (0.67)	1.44 (1.89)	0.82 (1.70)	1.10 (2.66)	0.84 (1.24)
14	cm.DESSC	Sentence count	11.67	4.28 (3.17)	6.36 (2.75)	7.49 (4.11)	10.17 (5.64)	14.29 (10.15)
15	lsa.similarity	Avg. LSA sim. between sentences	9.69	0.29 (0.27)	0.47 (0.23)	0.54 (0.23)	0.62 (0.20)	0.67 (0.17)
16	cm.DESSL	Avg. sentence length	9.60	11.88 (6.82)	13.62 (5.85)	16.69 (6.54)	19.36 (8.39)	21.73 (8.61)
17	cm.DESWLsyd	SD of word syllables count	8.92	0.98 (0.69)	1.33 (0.70)	0.98 (0.18)	0.97 (0.14)	0.97 (0.11)
18	liwc.i	Number of FPS* pronouns	8.84	4.33 (3.53)	2.82 (2.06)	2.37 (1.94)	2.51 (1.65)	2.19 (1.23)
19	cm.RDFKGL	Flesch-Kincaid Grade Level	8.29	7.68 (4.28)	10.30 (3.50)	10.19 (3.11)	11.13 (3.46)	11.99 (3.37)
20	cm.SMCAUSwn	WordNet overlap between verbs	8.14	0.38 (0.25)	0.48 (0.20)	0.51 (0.13)	0.50 (0.10)	0.47 (0.06)

MDG - Mean decrease Gini impurity index, **FPS** - first person singular

with only 205 classification features in total, which is $\sim$ 100x smaller than the number of bag-of-words features extracted by Kovanović et al. [35] classifier. This limits the chances of over-fitting the training data and also improves the performance of the classifier. This is particularly important for the prospective use of the classifier in different subject domains, and pedagogical contexts.

Another important finding of this study is the list of important classification features. We see that a small subset of features is highly predictive of the different phases of cognitive presence, while a majority of the features have a much lower predictive power (Figure 4). It is interesting to notice that most of the discussion context features (except the discussion start/end indicators) obtained high importance scores, indicating the value in providing contextual information to the classification algorithm. In our future work, we will focus on investigation of the additional features that would provide even more contextualized information to the classifier.

It is important to notice that the list of the most important variables is aligned with the conceptions of cognitive presence in the existing CoI literature. If we look at the messages in the four phases of cognitive presence, we can see that the higher levels of cognitive presence are associated with messages that are i) generally longer, with more sentences and paragraphs, ii) adopt more complex language with generally longer sentences, iii) include more named entities (e.g., names of different constructs, theories, people, companies, and geographical locations) iv) have lower lexical diversity, v) occur later in the discussion, vi) have higher givenness of the information, higher coherence, and higher verb overlap, vii) use fewer question marks and first-person singular pronouns, viii) exhibit higher similarity with the previous messages, and ix) more frequently use money-related terms. Interestingly, the feature of the highest importance is also the simple word count implying that the longer the message, the more likely it is in the higher levels of cognitive presence cycle. This is also consistent with the findings of a previous study with the same dataset [33]. Joksimović et al. [33] found that word count was the only LIWC 2007 variable that yielded statistically significant differences among all four cognitive presence categories. This is not totally surprising as the similar findings are reported by essay grading studies who found that the strongest predictor of the final essay grade is the length of the essay [48].

Looking at the non-cognitive or "other" messages, we can see that they are characterized by the large lexical diversity. This is expected, as non-cognitive messages tend to be shorter (i.e., fewer words, paragraphs, and sentences) and more informal. Higher levels of lexical diversity are known to be associated with very short tests or texts of low cohesion [10]. As "other" messages often are not related to the course topic, they also tend to have a lower number of named entities, and lower givenness and verb overlap. Such messages also tend to adopt a simpler language, as indicated by the lowest scores on the Flesch-Kincaid grade level. "Other" messages also tend to occur more frequently near the end of the discussion, as indicated by their high values for `message.depth` feature and also more often are related to expression of personal information, as indicated by the highest values for the use of first-person singular pronouns. This is expected as many discussions would typically finish with students thanking each other for their contributions.

6. CONCLUSIONS

This paper has twofold contributions. First, we developed a classifier for coding student discussion transcripts for the levels of cognitive presence with a much higher performance (0.63 Cohen's κ) than previously reported ones [35, 62] in the studies with the same dataset. The performance of the developed classifier is in the range which is generally considered to be a substantial level of agreement [39]. We can see that the proposed approach, which is based on the use of Coh-Metrix, LIWC, and discussion context features, shows a great promise for providing a fully automated system for coding cognitive presence. The feature space that is used is also much smaller, which limits the chances for over-fitting the data and makes the developed classifier more generalizable to other contexts.

Secondly, we can see a particular subset of classification features that are very highly predictive of the different phases of cognitive presence. The most predictive feature is simple word count, which implies that the longer the message is, the higher the chances are for the message to display higher levels of cognitive presence. We also identified several additional features which are also highly predictive of the cognitive presence phase, in particular the number of named entities that are used (higher values are associated with integration and resolution phase) and lexical diversity (lower values are associated with "other" and triggering messages). We also see that features that provide information on the discussion context

(i.e., similarity the with previous/next message, order in the discussion thread, and number of replies) are highly valuable and provide important information to the classification algorithm.

In our future work, we will focus on exploring additional features for improving the classification performance [43]. The study presented in this paper and our previous work [35] indicate that contextual features have a significant effect on classification accuracy and we will examine additional features of this kind. As our results reveal that the number of named entities has a significant effect on classification accuracy, and we will further explore similar features, such as concept maps [64], which would provide additional information about relationships between important concepts discussed in text-based messages. Finally, we will look at the different data preprocessing steps, including the use of the different algorithms for resolving the class imbalance problem. As we also observed that some of the students used direct quotes of other student messages which can cause problems for many of the text metrics that we used for classification, we will further examine the effects of the quotation on the final classification accuracy.

Finally, following the results presented in [17], we are exploring ideas for the development of a system that would – beside class labels – provide associated probabilities. Such a classifier could be used to develop a semi-automated classification system in which only one part of the data for which probabilities are sufficiently high would be automatically classified, and the rest would be manually classified. This would be advantageous as the *combined* desired accuracy of automatic-manual coding could be reached by setting a corresponding probability threshold. For achieving high levels of accuracy, a large majority of data would be classified automatically eliminating the large part of the manual work. Besides using it for coding discussion transcripts for research purposes, such system could be use, for example, to provide a real-time overview of the progress for a group of students and to point out the students for which an progress estimates are uncertain.

References

[1] Z. Akyol, J. B. Arbaugh, M. Cleveland-Innes, D. R. Garrison, P. Ice, J. C. Richardson, and K. Swan. A response to the review of the community of inquiry framework. *Journal of distance education*, 23(2), 2009. URL: http://www.ijede.ca/index.php/jde/article/view/630/884.

[2] T. Anderson and J. Dron. Three generations of distance education pedagogy. *The international review of research in open and distance learning*, 12(3):80–97, 2010. URL: http://www.irrodl.org/index.php/irrodl/article/view/890/.

[3] T. Anderson, L. Rourke, D. R. Garrison, and W. Archer. Assessing teaching presence in a computer conferencing context. *Journal of asynchronous learning networks*, 5:1–17, 2001. URL: http://auspace.athabascau.ca/handle/2149/725.

[4] J. B. Arbaugh, A. Bangert, and M. Cleveland-Innes. Subject matter effects and the community of inquiry (coi) framework: an exploratory study. *The internet and higher education*, 13(1):37–44, 2010. DOI: 10.1016/j.iheduc.2009.10.006.

[5] J. Arbaugh, M. Cleveland-Innes, S. R. Diaz, D. R. Garrison, P. Ice, J. C. Richardson, and K. P. Swan. Developing a community of inquiry instrument: testing a measure of the community of inquiry framework using a multi-institutional sample. *The internet and higher education*, 11(3–4):133–136, 2008. DOI: 10.1016/j.iheduc.2008.06.003.

[6] L. Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001. DOI: 10.1023/A:1010933404324.

[7] D. L. Butler and P. H. Winne. Feedback and self-regulated learning: a theoretical synthesis. *Review of educational research*, 65(3):245–281, 1995. DOI: 10.3102/00346543065003245.

[8] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004. DOI: 10.1145/1007730.1007733.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*:321–357, 2002. URL: https://www.jair.org/media/953/live-953-2037-jair.pdf.

[10] Coh-Metrix 3.0 indicies. http://cohmetrix.com/documentation_indices.html.

[11] S. Corich, K. Hunt, and L. Hunt. Computerised content analysis for measuring critical thinking within discussion forums. *Journal of e-learning and knowledge society*, 2(1), 2012. URL: http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/700.

[12] B. De Wever, T. Schellens, M. Valcke, and H. Van Keer. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: a review. *Computers & education*, 46(1):6–28, 2006. DOI: 10.1016/j.compedu.2005.04.005.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the american society for information science*, 41(6):391–407, 1990. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[14] J. Dewey. My pedagogical creed. *School journal*, 54(3):77–80, 1897.

[15] R. Donnelly and J. Gardner. Content analysis of computer conferencing transcripts. *Interactive learning environments*, 19(4):303–315, 2011. URL: http://eprints.teachingandlearning.ie/3930/.

[16] N. Dowell, O. Skrypnyk, S. Joksimović, A. C. Graesser, S. Dawson, D. Gašević, P. d. Vries, T. Hennis, and V. Kovanović. Modeling Learners' Social Centrality and Performance through Language and Discourse. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. Madrid, Spain, 2015. URL: http://www.educationaldatamining.org/EDM2015/proceedings/full250-257.pdf.

[17] P. Dönmez, C. Rosé, K. Stegmann, A. Weinberger, and F. Fischer. Supporting CSCL with automatic corpus analysis technology. In *Proceedings of th 2005 conference on computer support for collaborative learning: learning 2005: the next 10 years!*, 2005, 125–134. URL: https://telearn.archives-ouvertes.fr/hal-00190638.

[18] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014. URL: http://jmlr.org/papers/v15/delgado14a.html.

[19] P. Ferragina and U. Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *Software, ieee*, 29(1):70–75, 2012. DOI: 10.1109/MS.2011.122.

[20] P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25:285–307, 1998. URL: http://eric.ed.gov/?id=EJ589329.

[21] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, pp. 1606–1611. URL: http://dl.acm.org/citation.cfm?id=1625275.1625535.

[22] D. R. Garrison, T. Anderson, and W. Archer. Critical inquiry in a text-based environment: computer conferencing in higher education. *The internet and higher education*, 2(2–3):87–105, 1999. DOI: 10.1016/S1096-7516(00)00016-6.

[23] D. R. Garrison, T. Anderson, and W. Archer. Critical thinking, cognitive presence, and computer conferencing in distance education. *American journal of distance education*, 15(1):7–23, 2001. DOI: 10.1080/08923640109527071.

[24] D. R. Garrison, T. Anderson, and W. Archer. The first decade of the community of inquiry framework: a retrospective. *The internet and higher education*, 13(1–2):5–9, 2010. DOI: 10.1016/j.iheduc.2009.10.003.

[25] R. Garrison, M. Cleveland-Innes, and T. S. Fung. Exploring causal relationships among teaching, cognitive and social presence: student perceptions of the community of inquiry framework. *The internet and higher education*, 13(1–2):31–36, 2010. DOI: 10.1016/j.iheduc.2009.10.002.

[26] D. Gašević, O. Adesope, S. Joksimović, and V. Kovanović. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The internet and*

higher education, 24:53–65, 2015. DOI: 10.1016/j.iheduc.2014.09.006.

[27] L. Getoor. *Introduction to Statistical Relational Learning*. MIT Press, 2007. ISBN: 978-0-262-07288-5.

[28] P. Gorsky, A. Caspi, I. Blau, Y. Vine, and A. Billet. Toward a coi population parameter: the impact of unit (sentence vs. message) on the results of quantitative content analysis. *The international review of research in open and distributed learning*, 13(1):17–37, 2011. URL: http://www.irrodl.org/index.php/irrodl/article/view/1073.

[29] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich. Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational researcher*, 40(5):223–234, 2011. DOI: 10.3102/0013189X11413260.

[30] O. R. Holsti. *Content analysis for the social sciences and humanities*. Addison-Wesley Reading, MA, 1969.

[31] M. K. C. f. Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, t. R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, and C. Candan. *Caret: classification and regression training*. R package version 6.0-58, 2015. URL: http://CRAN.R-project.org/package=caret.

[32] S. Joksimović, N. Dowell, O. Skrypnyk, V. Kovanović, D. Gašević, S. Dawson, and A. C. Graesser. Exploring the Accumulation of Social Capital in cMOOC Through Language and Discourse. *Submitted*, 2015.

[33] S. Joksimović, D. Gašević, V. Kovanović, O. Adesope, and M. Hatala. Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. *The internet and higher education*, 22:1–10, 2014. DOI: 10.1016/j.iheduc.2014.03.001.

[34] S. Joksimović, V. Kovanović, J. Jovanović, A. Zouaq, D. Gašević, and M. Hatala. 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*. Poughkeepsie, NY, 2015, pp. 156–165. DOI: 10.1145/2723576.2723609.

[35] V. Kovanović, S. Joksimović, D. Gašević, and M. Hatala. Automated Content Analysis of 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. URL: http://ceur-ws.org/Vol-1137/.

[36] V. Kovanović, S. Joksimović, D. Gašević, M. Hatala, and G. Siemens. Content Analytics: the definition, scope, and an overview of published research. In, *Handbook of Learning Analyitcs*, 2015.

[37] K. H. Krippendorff. *Content analysis: an introduction to its methodology*. Sage Publications, 2003.

[38] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning (ICML '01)*, 2001. URL: http://dl.acm.org/citation.cfm?id=655813.

[39] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. DOI: 10.2307/2529310.

[40] A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. URL: http://CRAN.R-project.org/doc/Rnews/.

[41] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems 26*, 2013, pp. 431–439. URL: http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/281.pdf.

[42] R. Luppicini. Review of computer mediated communication research for education. *Instructional science*, 35(2):141–185, 2007. DOI: 10.1007/s11251-006-9001-6.

[43] E. Mayfield and C. Penstein-Rosé. Using feature construction to avoid large feature spaces in text classification. In *Proceedings of the 12th annual conference on genetic and evolutionary computation*, 2010, 1299–1306. DOI: 10.1145/1830483.1830714.

[44] T. McKlin. Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network. PhD thesis. Atlanta, GA, United States: Georgia State University, College of Education, 2004.

[45] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, 2014.

[46] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, 2011, 1–8. DOI: 10.1145/2063518.2063519.

[47] J. Mu, K. Stegmann, E. Mayfield, C. Rosé, and F. Fischer. The ACODEA framework: developing segmentation and classification schemes for fully automatic analysis of online discussions. *International journal of computer-supported collaborative learning*, 7(2):285–305, 2012. DOI: 10.1007/s11412-012-9147-y.

[48] E. B. Page and N. S. Petersen. The computer moves into essay grading: Updating the ancient test. *Phi delta kappan*, 76(7):561, 1995. URL: http://search.proquest.com/docview/218533317/abstract.

[49] C. L. Park. Replicating the Use of a Cognitive Presence Measurement Tool. *Journal of interactive online learning*, 8:140–155, 2, 2009. URL: http://www.ncolr.org/issues/jiol/v8/n2/replicating-the-use-of-a-cognitive-presence-measurement-tool#.VrVSebKUFhE.

[50] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer. Assessing social presence in asynchronous text-based computer conferencing. *The journal of distance education / revue de l'éducation à distance*, 14(2):50–71, 2007. URL: http://eric.ed.gov/?id=EJ616753.

[51] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer. Methodological issues in the content analysis of computer conference transcripts. *International journal of artificial intelligence in education (IJAIED)*, 12:8–22, 2001.

[52] P. J. Stone, D. C. Dunphy, and M. S. Smith. *The general inquirer: a computer approach to content analysis*. MIT press, 1966.

[53] J.-W. Strijbos. Assessment of (computer-supported) collaborative learning. *IEEE transactions on learning technologies*, 4(1):59–73, 2011. DOI: 10.1109/TLT.2010.37.

[54] J.-W. Strijbos, R. L. Martens, F. J. Prins, and W. M. G. Jochems. Content analysis: what are they talking about? *Computers & education*, 46(1):29–48, 2006. DOI: 10.1016/j.compedu.2005.04.002.

[55] M. Strube and S. P. Ponzetto. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*. AAAI Press, Boston, Massachusetts, 2006, pp. 1419–1424. ISBN: 978-1-57735-281-5. URL: http://dl.acm.org/citation.cfm?id=1597348.1597414.

[56] P.-N. Tan, V. Kumar, and M. Steinbach. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. ISBN: 0-321-32136-7.

[57] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1):24–54, 2010. DOI: 10.1177/0261927X09351676.

[58] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1):24–54, 2010. DOI: 10.1177/0261927X09351676.

[59] V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.

[60] J. Vassileva. Toward social learning environments. *IEEE transactions on learning technologies*, 1(4):199–214, 2008. DOI: 10.1109/TLT.2009.4.

[61] N. Vaughan and D. R. Garrison. Creating cognitive presence in a blended faculty development community. *The internet and higher education*, 8(1):1–12, 2005. DOI: 10.1016/j.iheduc.2004.11.001.

[62] Z. Waters, V. Kovanović, K. Kitto, and D. Gašević. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Proceedings of the 11th Asia Information Retrieval Societies Conference, AIRS 2015*, 2015. DOI: 10.1007/978-3-319-28940-3_18.

[63] I. H. Witten, E. Frank, and M. A. Hall. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 3rd ed., 2011.

[64] A. Zouaq and R. Nkambou. Building domain ontologies from text for educational purposes. *IEEE transactions on learning technologies*, 1(1):49–62, 2008. DOI: 10.1109/TLT.2008.12.