Dialogue attributes that inform depth and quality of participation in course discussion forums

Elaine Farrow
School of Informatics
University of Edinburgh
Edinburgh, UK
Elaine.Farrow@ed.ac.uk

Johanna Moore
School of Informatics
University of Edinburgh
Edinburgh, UK
J.Moore@ed.ac.uk

Dragan Gašević
Faculty of Information Technology
Monash University
Clayton 3800, Australia
Dragan.Gasevic@monash.edu

ABSTRACT

This paper describes work in progress to answer the question of how we can identify and model the depth and quality of student participation in class discussion forums using the content of the discussion forum messages. We look at two widely-studied frameworks for assessing critical discourse and cognitive engagement: the ICAP and Community of Inquiry (CoI) frameworks. Our goal is to discover where they agree and where they offer complementary perspectives on learning.

In this study, we train predictive classifiers for both frameworks on the same data set in order to discover which attributes are most predictive and how those correlate with the framework labels. We find that greater depth and quality of participation is associated with longer and more complex messages in both frameworks, and that the threaded reply structure matters more than temporal order. We find some important differences as well, particularly in the treatment of messages of affirmation.

CCS CONCEPTS

• **Computing methodologies → Model development and analysis**; *Supervised learning by classification*; • **Applied computing → Education**.

KEYWORDS

text analysis, discussion forum, participation, engagement, cognitive presence, Community of Inquiry, ICAP

ACM Reference Format:
Elaine Farrow, Johanna Moore, and Dragan Gašević. 2020. Dialogue attributes that inform depth and quality of participation in course discussion forums. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375462.3375481

1 INTRODUCTION

Discussion forums are widely used across all types of learning environments, from traditional face-to-face classroom settings to distance learning and MOOCs. It is increasingly common that the number of messages generated in a discussion forum is too large for instructors to monitor effectively. If the depth and quality of participation can be measured automatically while the course is still in progress, this could allow instructors to identify students who are bored, frustrated, or struggling, or lessons which cause confusion, while there is still time to intervene. For example, discussion forum transcripts could be colour-coded to indicate how the conversation was progressing, enabling instructors to see at a glance where to direct their attention.

Our aim in this initial study is to identify attributes of the dialogue that could be used in an automated system to discriminate between contributions of varying depth and quality, as measured by both the *phase of cognitive presence*, defined in the Community of Inquiry framework [4], and the *mode of cognitive engagement*, defined by the ICAP framework [2].

2 BACKGROUND

The Community of Inquiry (CoI) framework for online education is a powerful tool for analysing and developing effective learning experiences [4]. The framework identifies three main elements ('presences') that are important for a successful educational experience: a social environment conducive to learning (*social presence*), a well-designed course with ongoing facilitation (*teaching presence*), and the student's own cognitive engagement with the subject matter (*cognitive presence*). CoI has been widely used to analyse student learning in online courses, and predictive models have been developed for identifying its elements automatically using the text of discussion forum messages [3, 5, 6].

The ICAP framework [2] takes a different approach, defining cognitive engagement based purely on overt, observable, behaviours. The framework looks at individual learning activities and how they relate to students' cognitive engagement with the learning materials. Four 'modes', or levels, of engagement are identified, and the framework predicts that higher levels will be correlated with greater learning gains. The four levels, in descending order, are **I**nteractive, **C**onstructive, **A**ctive, and **P**assive. Each of these levels represents a qualitatively different *kind* of growth in knowledge, not simply a bigger or smaller change. Nevertheless, each level subsumes the levels below it. Off-task behaviours do not constitute engagement at any level. Prior work has demonstrated the feasibility of applying a modified version of the ICAP schema to MOOC discussion forums [12, 13] and to student comments on an annotated electronic course text [15] and MOOC videos [10].

While both frameworks address engagement, they do so from different perspectives. They were developed independently and

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375481

with different goals in mind. CoI was developed specifically in order to understand the benefit of online education and to explain how students develop their ideas through discussion leading to social knowledge construction. ICAP has a broader scope and has been demonstrated to be effective in predicting the educational value of several different interventions, in a classroom setting as well as online. Finding the commonalities and differences in how these frameworks apply to one specific data set offers a useful contribution to the theoretical understanding of online learning and learning through discussion. If the frameworks are found to be closely correlated, then results derived using each of them in previous studies can be expected to be applicable to work using the other. If instead they are completely distinct, then using them together in future studies will give a richer picture of student participation. Our expectation is that, while there will be some similarities, overall the frameworks will provide complementary views on learning.

3 METHODOLOGY

In order to address our research question, we used a data set of course discussion forum messages that was annotated with labels assigned by the two frameworks we are examining. We trained several random forest models and used the best of these to assign labels to the messages in a held-out test set. Having determined that the predictive performance was sufficiently good, we examined which of the dialogue attributes used as model features could discriminate between messages in terms of depth and quality, as measured by the outcome variables.

3.1 Description of the data

This work makes use of a data set that has previously been used in several studies of cognitive presence. It was collected from a fully online distance-learning course at a Canadian university that formed part of a Masters degree in software engineering. We use data from the first four course offerings, which took place in 2008 and 2009. The distribution of students and messages across the sessions is shown in Table 1.

Table 1: Statistics for the 4 course offerings used in this work.

Session	Student count	Message count
Winter 2008	16	212
Fall 2008	24	633
Spring 2009	12	243
Fall 2009	9	63
Average (SD)	15.3 (6.5)	287.8 (243.2)
Median	14	227.5
Total	61	1151

Each student created and shared a video presentation based on a research paper relevant to the course, then started a new thread in the discussion forum to host a conversation around their presentation. We do not have access to the presentations themselves, only to the text-based discussion that followed. Students were in general highly motivated since forum participation accounted for 10% of the final course mark.

3.2 Labels assigned by the frameworks

The messages in our data set had previously been annotated with their phase of cognitive presence by two expert coders (98.1% agreement, Cohen's $\kappa = 0.974$). Sometimes a message can show indications of two distinct phases of cognitive presence. The coding scheme indicates that these should be coded with the higher phase [14]. This is sometimes referred to as *coding up*. Table 2 shows the distribution of the CoI phases of cognitive presence across the data.

Table 2: Messages by CoI phases of cognitive presence.

Cognitive presence phase	Count	Percentage
Other	83	7.21%
Triggering Event	227	19.72%
Exploration	480	41.70%
Integration	293	25.46%
Resolution	68	5.91%
All	1151	100.00%

For this study, we additionally annotated each message with a label indicating the relevant cognitive engagement mode from the ICAP framework. In the labelling task itself, we built on earlier work [12, 13, 15] that developed guidelines for applying ICAP labels to data from MOOC discussions and annotated course texts using an extended label set that allows for finer-grained distinctions between messages within two of the modes: *Constructive* mode is divided into *Constructive Reasoning* and *Constructive Extending*, while *Active* mode is divided into *Active Targeted* and *Active General*.

In the prior work, affirmation messages consisting primarily of agreement or thanks expressed in response to an earlier message were treated as a special case: the label assigned to them depended on the label of that earlier message. If the earlier message was labelled as *Interactive* or *Constructive Reasoning*, then the affirmation message was labelled as *Constructive Extending*; in all other cases, the affirmation message simply inherited the earlier label.

However, for the purpose of developing an automated classifier that can label future data reliably, it is preferable to assign each label based only on attributes of the current message. Otherwise, two affirmation messages with identical content (*e.g.*, "Thanks for your reply") and appearing in the same position within a thread could receive different labels depending on the labels of the earlier messages. Therefore, in the current work, we do not assign the derived label to affirmation messages directly. Instead, we give them the *Affirmation* label as a placeholder. Once all the messages in the data set have had labels assigned (by manual coding or using an automated classifier), a simple rule-based transformation can be applied to relabel all *Affirmation* messages, based on the labels that were assigned to the messages they are affirming.

Each message was assigned a single label, corresponding to the highest level of cognitive engagement that was identified in the message, similar to the *coding up* process that was used for the CoI labels. The full extended ICAP label set we used for annotation, along with the distribution of the labels across the data set, is presented in Table 3. We have no access to data indicating when a student read a message without responding, so the *Passive* label is not used in this study.

Table 3: Breakdown of messages in the current data set across modes in the extended cognitive engagement taxonomy, adapted from Yogev et al. [15] and based on the ICAP framework.

Cognitive engagement mode	Example behaviour	Count	Percentage
Interactive	Displaying explanation or reasoning about the current topic in response to an earlier message	373	32.41%
Constructive Reasoning	Displaying explanation or reasoning about the current topic	187	16.25%
Constructive Extending	Introducing new content to the discussion	296	25.72%
Active Targeted	Referencing specific previous content	180	15.64%
Active General	Showing other signs of being engaged with course content	61	5.30%
Passive	Reading messages without responding	0	0.00%
Affirmation	Affirming what was said in an earlier message	53	4.60%
Off-task	Commenting without any relation to the current topic or the course	1	0.09%
All		1151	100.00%

3.3 Dialogue attributes used as model features

Each message was annotated with 206 classification features: 91 word counts derived using the LIWC software package [11], 106 metrics related to text coherence, complexity, readability, and lexical category use from Coh-Metrix [9], and 9 features capturing aspects of the discussion structure, described next and shown in Table 4. The Coh-Metrix and LIWC features are the same as those used in earlier work using this data [3, 6].

The threaded nature of the forum means that every message can receive multiple replies, and replies can themselves receive replies. A new reply can be added at any level in the chain at any time. Without knowing which messages a student has actually read, we need to make some assumptions. A message posted as a reply to another message can be expected to relate to that message in a meaningful way. Similarly, the impact of a message on the discussion can be measured not only by the number of replies it gets, but perhaps also by the total count of replies-to-replies: that is, counting all the descendant messages. We thus defined three features related to message position in the thread (message depth, first message, and last message) and two features for replies (number of direct replies and total number of replies).

We expected that the chronological order of messages would also be relevant, so we ordered the messages within each thread using time-stamp order and then derived features using that ordering: (position from start, position from end, and fractional position).

A final feature (discussion size) captures the total number of messages in the thread, allowing the classifier to distinguish between longer and shorter discussions.

Table 4: Structural features derived from the data and used as model features.

Structural feature	Description and Rationale
discussion size	The total number of messages in the current discussion. We speculate that a short discussion is less likely to progress to deeper phases of cognitive presence than a longer one.
position from start	The index of the message in chronological order from the beginning of the discussion. Early messages may be more likely to introduce new material.
position from end	The index of the message in chronological order from the end of the discussion. Later messages may build on what has gone before to achieve greater depth and quality.
fractional position	The position of the message chronologically within the discussion, as a fraction of the total discussion size. This feature seeks to allow for natural variations in discussion length.
number of direct replies	The number of direct replies to the message. Messages relating to *triggering event*s and *exploration* are expected to generate more replies than those in deeper phases [14].
total number of replies	The cumulative number of direct and indirect replies (replies to replies). Exchanges with a partner are a key feature of the *Interactive* mode. This feature may also capture the role of *triggering event*s and *exploration* better than direct replies alone [8].
depth in thread	The depth of the message within the threaded view of the discussion. A deeper message is more likely to be an example of *Interactive* engagement.
first/ last message	Binary indicators for the first and last message in a discussion thread, defined chronologically.

3.4 Method

We used the first three offerings of the course as training data for a random forest classifier, and kept back the data from the fourth session as unseen test data with which to assess the best model. For this initial study, we recombined the finer-grained distinctions within the *Constructive* and *Active* modes, in common with prior work [12, 15]. As there were so few *Off-task* messages, those records were excluded from our analysis of the ICAP framework. We explored 20 different settings for the `mtry` parameter that controls how many of the 206 classification features are available as candidates at each split point. The specific values to be tested are automatically determined by the `caret` library in `R` based on the number of features in the model; here, they were 2, 12, 23, 34, 44, 55, 66, 77, 87, 98, 109, 120, 130, 141, 152, 163, 173, 184, 195, and 206. For each `mtry` setting, we trained 1,000 trees and used 10-fold cross-validation, repeated 10 times, to select the best performing value. A final random forest model was built using this value and data from the full training set.

The number of data points belonging to each outcome class (*i.e.* the phases of cognitive presence and the ICAP modes) is unbalanced (Tables 2 and 3) and it is well-known that unbalanced data can cause problems for classification techniques. For this reason, we also compared models trained directly on the unbalanced training data against models using SMOTE (Synthetic Minority Over-sampling TEchnique [1]) to rebalance the classes in the outcome variable such that every outcome class had the same size. Following best practice, the SMOTE algorithm is run inside the cross-validation loop so that the class rebalancing step for each fold of the cross-validation uses only the training data for that fold, avoiding a potential source of data contamination [3].

For each framework, the model that achieved the highest Cohen's κ score in cross-validation was used to assign labels to the held-out test data, and the relative importance of each variable in the model was compared. In this way, we identified the dialogue attributes that were most predictive of the different CoI phases of cognitive presence and ICAP modes of cognitive engagement.

4 RESULTS AND ANALYSIS

4.1 Predictive performance metrics

When dealing with unbalanced classes, as we are here, Cohen's κ and the macro-averaged F_1 score are more informative than accuracy. We chose the best model for each framework based on Cohen's κ (Table 5). In each case, rebalancing the outcome classes using SMOTE inside the cross-validation loop gave better results during training than using the original unbalanced data.

Table 5: Cross-validation results: outcome metrics and the best value for the `mtry` tuning parameter, with and without class rebalancing using SMOTE.

Pre-processing	Cohen's κ	Macro F_1	Best `mtry`
CoI phases of cognitive presence			
None	0.411	0.492	44
Rebalance with SMOTE	**0.421**	**0.539**	34
ICAP modes of cognitive engagement			
None	0.581	0.667	44
Rebalance with SMOTE	**0.592**	**0.678**	55

We used the best model from each framework to assign labels to the held-out data from the fourth offering of the course. The Cohen's κ scores are shown in Table 6, along with the Precision, Recall, and F_1 scores for each class of the outcome variable. We found that Cohen's κ scores were higher for both frameworks on the test data than the estimates from cross-validation. For the model based on CoI phases of cognitive presence, a Cohen's κ of 0.428 indicates a 'moderate' level of agreement with the gold-standard human coding, while the Cohen's κ of 0.645 for the ICAP modes of cognitive engagement indicates 'substantial' agreement [7]. The macro-averaged F_1 score for the CoI model was 0.570, again demonstrating an improvement over the cross-validation estimate. As there were in fact no *Affirmation* messages in the held-out data, the macro-averaged F_1 score for the ICAP model is not meaningful.

Table 6: Outcome metrics on the held-out test data

	Precision	Recall	F_1	Cohen's κ
CoI phases of cognitive presence				0.428
Other	1.000	0.333	0.500	
Triggering Event	0.800	0.800	0.800	
Exploration	0.485	0.800	0.604	
Integration	0.750	0.429	0.546	
Resolution	0.333	0.500	0.400	
ICAP modes of cognitive engagement				0.645
Interactive	0.950	0.864	0.905	
Constructive	0.813	0.867	0.839	
Active	0.500	0.364	0.421	
Affirmation	0.000	NA	NA	

4.2 Analysis of variable importance

Looking more closely at the best models we see that, in each case, a small subset of features have a high degree of explanatory power, evidenced by their high Mean Decrease Gini (MDG) values (Figure 1). The top 20 features by importance are listed in Tables 7 and 8 along with their mean values for each class of the outcome variable.

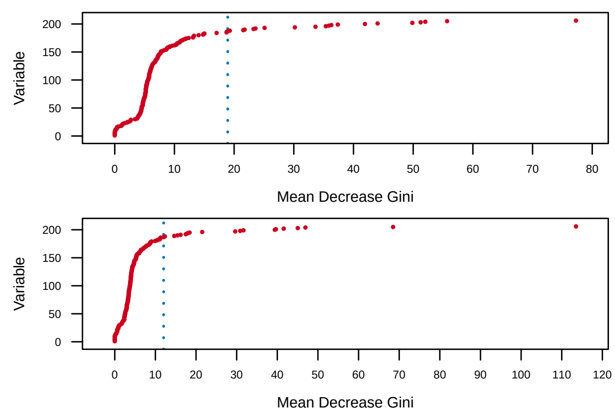

Figure 1: Mean Decrease Gini indicating variable importance in the models: (top) CoI phases of cognitive presence; (bottom) ICAP modes of cognitive engagement. In each case, the vertical dotted line separates the top 20 features.

Longer messages and fewer question marks were associated with deeper engagement in both frameworks: the number of words in the message, the number of sentences, and the mean paragraph length all appear in the top 20 for both models. Similarly, lower levels of lexical diversity (measured by type-token ratio) were associated with deeper phases of cognitive presence and also with deeper cognitive engagement. In contrast, when using the alternative VOCD lexical diversity metric that aims to compare texts of different lengths more reliably, the relationship was reversed: *higher* levels of lexical diversity were seen to be associated with both deeper cognitive presence and cognitive engagement. These results are in line with prior work on CoI [6]. It is interesting to see that they apply to ICAP as well.

Table 7: The 20 most important variables in the CoI cognitive presence model, ranked by Mean Decrease Gini (MDG) and shown with their mean scores and standard deviations for messages in different phases of cognitive presence.

#	Variable	Description	MDG	Phase of cognitive presence				
				Other	Triggering Event	Exploration	Integration	Resolution
1	cm.DESWC	Number of words	77.27	58.88 (92.42)	83.63 (44.08)	119.83 (72.50)	186.06 (105.30)	293.47 (247.86)
2	cm.WRDHYPn	Hypernyms for nouns	55.68	4.77 (1.64)	5.49 (0.93)	6.05 (0.69)	6.19 (0.58)	6.34 (0.58)
3	liwc.posemo	Number of expressions of +ve emotion	52.03	11.50 (14.78)	4.51 (2.81)	4.01 (2.19)	3.57 (1.72)	2.97 (1.48)
4	liwc.discrep	Number of discrepancy words	51.25	1.19 (2.01)	1.04 (1.43)	1.80 (1.66)	1.91 (1.39)	2.78 (1.57)
5	message.is.first	First message	49.86	1.00 (0.00)	1.25 (0.43)	1.00 (0.00)	1.00 (0.06)	1.00 (0.00)
6	message.depth	Message depth in discussion	44.04	2.69 (1.21)	1.06 (0.93)	1.84 (0.94)	1.87 (0.82)	2.06 (0.90)
7	cm.LDTTRa	Lexical diversity, all words	41.91	0.87 (0.13)	0.76 (0.08)	0.71 (0.10)	0.65 (0.09)	0.58 (0.10)
8	cm.WRDMEAc	Meaningfulness	37.37	348.03 (147.67)	406.72 (35.79)	404.76 (22.45)	405.32 (16.26)	406.05 (12.81)
9	liwc.affect	Number of affective process words	36.29	11.90 (14.62)	5.11 (2.97)	4.63 (2.43)	4.41 (2.06)	3.83 (1.65)
10	message.replies.all	Total number of replies	35.87	0.60 (1.56)	5.82 (8.85)	0.80 (1.06)	0.83 (1.14)	0.50 (0.87)
11	liwc.QMark	Number of question marks	35.36	0.25 (0.76)	1.81 (1.60)	0.82 (1.15)	0.52 (0.76)	0.39 (0.62)
12	message.replies.direct	Number of direct replies	33.63	0.37 (0.84)	2.69 (3.36)	0.56 (0.56)	0.55 (0.59)	0.35 (0.48)
13	cm.LSASSpd	SD of LSA overlap in paragraph	30.17	0.06 (0.06)	0.11 (0.06)	0.13 (0.06)	0.13 (0.04)	0.13 (0.03)
14	liwc.SemiC	Number of semi-colons	25.11	0.54 (2.51)	0.19 (0.69)	0.12 (0.41)	0.09 (0.28)	0.12 (0.36)
15	message.thread.size	Discussion size	23.58	22.77 (6.97)	24.37 (7.10)	24.47 (7.19)	22.17 (7.47)	22.97 (6.87)
16	liwc.money	Number of money words	23.25	0.23 (0.80)	0.30 (0.71)	0.38 (0.83)	0.61 (1.05)	0.89 (1.12)
17	cm.LDVOCD	Lexical diversity, VOCD	21.76	17.88 (43.63)	30.29 (49.57)	52.54 (54.03)	84.52 (44.92)	95.10 (37.67)
18	cm.DESPL	Mean length of paragraphs	21.53	4.47 (4.23)	6.28 (2.85)	7.41 (4.34)	10.38 (5.77)	15.31 (13.74)
19	cm.DESSC	Number of sentences	19.27	4.47 (4.23)	6.28 (2.85)	7.41 (4.34)	10.38 (5.77)	15.31 (13.74)
20	cm.DESWLltd	SD of word length in letters	18.93	3.18 (2.24)	3.86 (1.92)	2.90 (0.40)	2.90 (0.43)	2.85 (0.28)

Table 8: The 20 most important variables in the ICAP model, ranked by Mean Decrease Gini (MDG) and shown with their mean scores and standard deviations for messages in different modes of cognitive engagement.

#	Variable	Description	MDG	Mode of cognitive engagement			
				Interactive	Constructive	Active	*Affirmation*
1	message.depth	Message depth in discussion	113.50	2.30 (0.68)	1.34 (0.94)	1.58 (1.02)	2.79 (1.10)
2	liwc.assent	Number of expressions of assent	68.50	0.27 (0.76)	0.24 (1.22)	0.11 (0.35)	1.40 (2.33)
3	liwc.posemo	Number of expressions of +ve emotion	46.91	3.33 (1.89)	3.82 (2.22)	5.51 (3.89)	13.85 (16.78)
4	message.replies.all	Total number of replies	45.04	0.42 (0.89)	3.24 (6.52)	1.24 (1.84)	0.25 (0.48)
5	message.replies.direct	Number of direct replies	41.58	0.25 (0.46)	1.60 (2.54)	0.88 (0.70)	0.23 (0.42)
6	cm.DESWC	Number of words	39.66	191.94 (152.04)	125.26 (79.38)	87.98 (55.17)	47.40 (44.90)
7	cm.LSAGN	LSA Given-New	39.38	0.21 (0.07)	0.17 (0.07)	0.15 (0.06)	0.08 (0.07)
8	liwc.QMark	Number of question marks	31.68	0.24 (0.57)	1.02 (1.20)	1.70 (1.50)	0.22 (0.66)
9	liwc.affect	Number of affective process words	30.88	4.11 (2.19)	4.43 (2.44)	6.18 (3.89)	14.32 (16.57)
10	cm.LDTTRa	Lexical diversity, all words	29.65	0.65 (0.11)	0.71 (0.10)	0.76 (0.10)	0.87 (0.12)
11	cm.LDVOCD	Lexical diversity, VOCD	21.52	73.66 (46.92)	62.12 (57.66)	27.56 (45.47)	12.93 (35.18)
12	cm.DESPL	Mean length of paragraphs	18.41	10.32 (8.03)	8.06 (4.85)	6.15 (3.24)	4.00 (2.56)
13	cm.RDFKGL	Flesch-Kincaid Grade Level score	17.98	11.22 (3.61)	10.98 (3.19)	9.12 (2.96)	6.51 (3.74)
14	liwc.you	Number of 2nd person personal pronouns	17.72	1.28 (1.26)	1.64 (1.49)	3.47 (3.36)	3.70 (5.24)
15	cm.DESSC	Number of sentences	17.48	10.32 (8.03)	8.06 (4.85)	6.15 (3.24)	4.00 (2.56)
16	cm.WRDPRP2	Number of 2nd person pronouns	16.22	12.69 (12.45)	16.39 (14.68)	34.42 (33.60)	35.11 (52.01)
17	liwc.Period	Number of periods	15.44	5.46 (1.81)	6.66 (4.27)	5.38 (2.51)	11.97 (10.92)
18	cm.RDFRE	Flesch Reading Ease score	14.64	49.67 (13.61)	46.03 (17.52)	55.92 (14.02)	66.84 (18.76)
19	liwc.ppron	Number of personal pronouns	12.34	4.68 (2.57)	4.85 (2.71)	7.40 (3.92)	8.36 (7.27)
20	liwc.tentat	Number of tentative words	12.03	3.74 (2.09)	3.26 (2.20)	3.61 (2.70)	2.39 (2.47)

Messages that were more deeply nested in the discussion (though not necessarily posted later in time) were more likely to come from the *other* or *resolution* phases of cognitive presence, and to indicate *Affirmation* or the *Interactive* mode of ICAP.

None of the features relating to time-stamp order within a thread appeared in the top 20 for ICAP, and the only one to do so for CoI was the 'first message' indicator, which is strongly indicative of *triggering event*s. However, the number of replies (both direct and indirect) a message received was highly predictive for both frameworks. The values were highest for messages labelled *Constructive* and *triggering event* respectively. Both of these observations can be explained by noting that the way the original discussion task was structured meant that the first message in each thread was nearly always assigned the same label: *triggering event* or *Constructive* (specifically, *Constructive Extending*). However, in the ICAP framework, the *Constructive* label is often used for follow-up messages, while in the CoI framework it is rare for another *triggering event* to be seen in the same thread before the first one has been resolved.

The number of expressions of positive emotion was the third most predictive feature in both models and the number of affective process words also appears in the top 10 in both lists. Both are strongly indicative of *other* messages (those that display no signs of cognitive presence) and of *Affirmation* messages.

Considering features that are predictive for one of the two frameworks but not the other, we see that messages displaying deeper levels of the CoI phases of cognitive presence used more words from the LIWC categories relating to discrepancies (such as *should* and *would*) and money (such as *owe*). Meanwhile, the Coh-Metrix measure tracking the amount of 'given' versus 'new' information in each sentence was highly predictive for the ICAP modes of cognitive engagement. The highest values are seen for *Interactive* messages, which are expected to build on and develop the arguments from earlier messages; and lowest for *Affirmation* messages. We also note that use of second person pronouns is strongly indicative of *Active* mode, where quoting is expected, and *Affirmation* messages.

4.3 Discussion

We observed many similarities between the predictors for the two frameworks we are investigating. Some are unsurprising: longer messages are correlated with deeper levels of engagement in both. Others are more complex. Messages displaying higher than average numbers of affective process words and expressions of positive emotion tend to cluster in a single class of the outcome variable (*other* and *Affirmation*, respectively). However, there are important differences in the interpretation of these classes.

Whereas the *other* label indicates that no signs of cognitive presence were evident in a given message, messages with the *Affirmation* label are later relabelled based on the label of the message to which they were responding. By affirming what was said in an earlier message, the student is thus credited with demonstrating some cognitive engagement, albeit not to the same extent as the original contributor (see the description of the relabelling process in Section 3.2 for details). Since the *Interactive* mode is associated with the greatest learning gains, ICAP rewards conversational moves that foster interactivity by continuing the conversation and opening the way for further elaboration. In contrast, the CoI framework treats messages of affirmation solely as indicators of social presence.

4.4 Limitations

Only a single data set was used for this preliminary study. Because of the particular discussion task that was set in that course, the first message of every thread follows a similar format and is typically labelled in the same way. There is no reason to suppose that messages from another course would share this property, so some caution is needed in interpreting results relating to features derived from message position. Additionally, the *Passive* mode of the ICAP framework was not used at all, because the data set does not include a record of when students read the messages posted by others, and *Off-task* messages were too infrequent to be used in this study.

5 CONCLUSION AND FUTURE WORK

Our aim was to identify dialogue attributes that could be used to discriminate between contributions of varying depth and quality. Our expectation was that the CoI and ICAP frameworks would provide complementary perspectives. We found that several simple measures of contribution size (such as the number of words and sentences) correlated with deeper engagement in both frameworks; while other correlations were framework-specific, such as the higher numbers of second-person pronouns found in the *Active* mode in ICAP. The reply-based network structure in the message threads proved to be more important than chronological order, particularly in the ICAP model, and so we recommend that users and providers of discussion boards should ensure that such information is always preserved and made available for analysis.

We also considered the different treatment of affirmations in the two frameworks. In CoI, they are considered solely as indicators of social presence, with no value in terms of cognitive presence; whereas with ICAP their value depends on the content of the earlier message they are affirming, due to the greater value placed on interactivity. While contribution quantity is highly correlated with measures of participation, simply setting a minimum threshold on message length is unlikely to improve learning and would certainly harm social exchanges such as affirmations. Future research should look beyond contribution quantity to consider which other dialogue attributes indicate quality of participation.

In our own future work we will further investigate the relationship between the labels assigned by the two frameworks examined here, using visualisation techniques and network-based analysis, as well as looking at co-occurrence metrics.

ACKNOWLEDGMENTS

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

REFERENCES

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. https://doi.org/10.1613/jair.953

[2] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. https://doi.org/10.1080/00461520.2014.965823

[3] Elaine Farrow, Johanna Moore, and Dragan Gašević. 2019. Analysing Discussion Forum Data: A Replication Study Avoiding Data Contamination. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19)*. ACM, New York, NY, USA, 170–179. https://doi.org/10.1145/3303772.3303779

[4] D. Randy Garrison, Terry Anderson, and Walter Archer. 1999. Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The Internet and Higher Education* 2, 2 (1999), 87–105. https://doi.org/10.1016/S1096-7516(00)00016-6

[5] Dragan Gašević, Olusola Adesope, Srećko Joksimović, and Vitomir Kovanović. 2015. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The Internet and Higher Education* 24 (2015), 53–65. https://doi.org/10.1016/j.iheduc.2014.09.006

[6] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)*. ACM, New York, NY, USA, 15–24. https://doi.org/10.1145/2883851.2883950

[7] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[8] Thomas Edward McKlin. 2004. *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Ph.D. Dissertation. Georgia State University.

[9] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, New York. https://doi.org/10.1017/CBO9780511894664

[10] Yassin Taskin, Tobias Hecking, H. Ulrich Hoppe, Vania Dimitrova, and Antonija Mitrovic. 2019. Characterizing Comment Types and Levels of Engagement in Video-Based Learning as a Basis for Adaptive Nudging. In *Proceedings of the 14th European Conference on Technology Enhanced Learning*. Springer International Publishing, Cham, 362–376.

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

[12] Xu Wang, Miaomiao Wen, and Carolyn P Rosé. 2016. Towards Triggering Higher-order Thinking Behaviors in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)*. ACM, New York, NY, USA, 398–407. https://doi.org/10.1145/2883851.2883964

[13] Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé. 2015. Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society, Massachusetts, USA, 226–233.

[14] Zak Waters, Vitomir Kovanović, Kirsty Kitto, and Dragan Gašević. 2015. Structure Matters: Adoption of Structured Classification Approach in the Context of Cognitive Presence Classification. In *Information Retrieval Technology*, Vol. 9460. Springer International Publishing, Cham, 227–238. https://doi.org/10.1007/978-3-319-28940-3_18

[15] Eran Yogev, Kobi Gal, David Karger, Marc T. Facciotti, and Michele Igo. 2018. Classifying and Visualizing Students' Cognitive Engagement in Course Readings. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale (L@S '18)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/3231644.3231648