

# Can Online Discussion Participation Predict Group Project Performance? Investigating the Roles of Linguistic Features and Participation Patterns

Jaebong Yoo · Jihie Kim

Published online: 22 October 2013

© International Artificial Intelligence in Education Society 2013

**Abstract** Although many college courses adopt online tools such as Q&A online discussion boards, there is no easy way to measure or evaluate their effect on learning. As a part of supporting instructional assessment of online discussions, we investigate a predictive relation between characteristics of discussion contributions and student performance. Inspired by existing work on dialogue acts, project-based learning, and instructional analysis of student-generated text in generating predictive models, we make use of dialogue roles, linguistic features, and work patterns. In particular, we model the Q&A dialog roles that participants play, emotional features covered by LIWC (Linguistic Inquiry and Word Count), cohesiveness of the dialogue, the coherence captured by Coh-Metrix, and temporal patterns of participation. We use a discussion corpus from eight semesters of a computer science course, covering conversations of 173 student groups (370 students). We first remove various noises in student discussion data and normalize the discussion data. We then apply machine learning techniques and text analysis tools for classifying dialogue features efficiently. The extracted dialogue and participation features are used as predictive variables for project grades. The correlation and regression analyses indicate that the number of answers provided to others, the number of positive emotion expressions, and how early students communicate their problems before the deadline correlate with project grades. This finding confirms the argument that in assessing student online activities, we need to capture how they interact, not just how often they participate.

**Keywords** Student online discussions · Performance prediction · Dialog analysis · Computational linguistic features · Work patterns · Educational data mining

---

J. Yoo (✉) · J. Kim

University of Southern California/Information Sciences Institute, 4676 Admiralty Way,  
Marina del Rey, CA 90292, USA  
e-mail: jaebong@isi.edu

J. Kim

e-mail: jihie@isi.edu

## Introduction

Online discussion boards, or online asynchronous discussions, have become increasingly popular tools for university-level engineering courses in supporting students' communication and collaboration. Integrated into most virtual learning environments, online discussion boards allow students and instructors to exchange ideas and information, often in a form of Q&A threads.

Some studies have pointed to online discussion boards as a promising strategy for collaborative problem solving and higher-order thinking (Scardamalia and Bereiter 1994; Koschmann 1996; Harman and Koochang 2005; Erlin et al. 2009). First, there are significant learning benefits to collaborative learning, inquiry and guided discovery-oriented activities, and reflective interaction amongst students (Bransford 2000; Felder and Henriques 1995). With respect to online learning, research indicates that barriers to effective discussion participation in face-to-face interaction may be ameliorated in online interactions and students may have greater autonomy (Steele 1997; Osborne 2003; Ambady et al. 2001; Krupnick 1985).

In order to assess the effectiveness of online discussions, researchers have sought approaches for evaluating student discussion activities. A large body of research applies quantitative and qualitative methods. Quantitative approaches often use statistical information such as message frequency, the number of initial messages and replies, view counts, thread lengths, and response time for the given message (Palmer et al. 2007; Davies and Graff 2005; Kay 2006). Qualitative approaches that rely on characteristics of message content have gained considerable attention in the past decade (De Wever et al. 2006). Such approaches can reveal latent semantic information in transcripts from discussion boards on student knowledge building (Gunawardena et al. 1997) and critical thinking behavior (Newman et al. 1997). Some researchers have applied machine learning techniques for efficient analysis of student contributions (McLaren et al. 2007). However, such results have not been fully used for explaining or predicting student performance in the class. For example, in project-based engineering courses, participation in discussions on class projects may facilitate better project work. In addition, discussion content and participation patterns can reveal characteristics of students' contributions and work patterns. The main objective of our work is to evaluate the hypothesis that discussion participation patterns provide good indicators for related project performance. Our work builds on existing research in relevant areas including pedagogical analysis of student text, project-based learning, and evaluation of online discussions.

Past studies of online asynchronous discussions analyzed characteristics of participation including the degree of participation (De Wever et al. 2006). Assessment of student written text made use of types of words, such as complex words, used by students (McNamara et al. 2010), and emotional expressions in the text (D'Mello and Graesser 2012). Collaborative dialogue behavior was analyzed in the context of peer tutoring or tutor/tutee interactions (Kumar and Rose 2011) and in the process of deliberation (Murray et al. 2012). Types of contributions made by students or roles that students played were analyzed for discovering student difficulties (Ravi and Kim 2007). Research on student group projects investigated work patterns of the participants such as work pacing in performing the project (Ganapathy et al. 2011). These results are useful for evaluating student written texts or collaborative dialogues

including interactions with intelligent tutoring systems. However, they have not been fully used in analyzing project-based online discussions. We investigate student participation in such discussions and how the discussion data can be used for creating a predictive model of student performance. That is, we are interested in using online discussion participation as an indicator for their project work and building a tool for predicting student project performance. In particular, we aim to evaluate the role of dialogue and interaction features that have been useful for assessing student written text and collaborative dialogues. The following research questions have been formulated based on related past studies:

- What is the relationship between the degree of participation or quantity of messages posted and the project performance? E.g., if there are more messages with more words written, the student may have been more engaged in the information exchange and have more opportunities to reflect on the project related problems.
- What is the relationship between the kinds of words used by the student or coherency of sentences and the project performance? E.g., if the student uses more complex words or coherent words in the message, the student may have better understanding of the problem or the subject than the other students who used simpler words or less coherent sentences.
- What is the relationship between the expressions of emotion and the project performance? E.g., if the student uses more positive expressions than others, he/she may have a more positive attitude toward the current topic, and may be more interested in the project.
- What is the relationship between the style of participation in the discussion, such as information seeking vs. giving, and the project performance? E.g., if the student provides more information to others, he/she may have a better understanding of the subject than the students who seek information from others.
- What is the relationship between work pacing and work performance? E.g., if the group works continuously throughout the project period, its members may do better than other groups who work only a few days before the deadline.

In addressing these questions, we make use of a relatively large corpus of student discussion contributions in computer science courses, and relate dialogue features that capture the dynamics of Q&A conversation to student performance. Generating meaningful features from student discussions is often challenging. The data includes a lot of noise including spelling mistakes, programming code, and quoted text. There also are many ways that students present similar information. The data must be properly cleaned and variances should be minimized before the analysis. We describe several approaches for data preparation and processing.

In analyzing characteristics of the words used or linguistic features, we make use of Coh-Metrix (Graesser et al. 2004), a computational tool that produces diverse measures for evaluating student written text. In particular, Coh-Metrix can capture cohesion, coherence, and readability of student-written text using linguistic and semantic characteristics of the words.

For evaluating affective expressions in online discussions, Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001) is used. LIWC is a validated measure that has been used in analyzing emotional and psychological features in diverse corpora of texts including texts in blogs and Twitter messages (Pennebaker et al.

2007; Quercia et al. 2011). For example, students who have difficulty solving the problems or who share unresolved issues close to the deadline might express negative emotions like sadness or frustration. On the other hand, students may present positive expressions such as thanking if they get answers or hints for a problem within a reasonable wait time.

For analyzing question and answer, or give-and-take relations among the students, we model student conversations using the Speech Acts (SAs) framework developed for dialogue analysis (Searle et al. 1980). We present a novel dialogue feature that captures the roles that individual students play in Q&A information exchange forums: *Sink* (information seeking act) and *Source* (information providing act). We provide a robust machine classification approach for automating such analysis using natural language processing and machine learning techniques.

We also propose a novel way of analyzing work pacing through a temporal analysis of discussion participation. In particular, we evaluate how early the students start asking questions or participating in discussions before the project deadline as an indicator for work pacing.

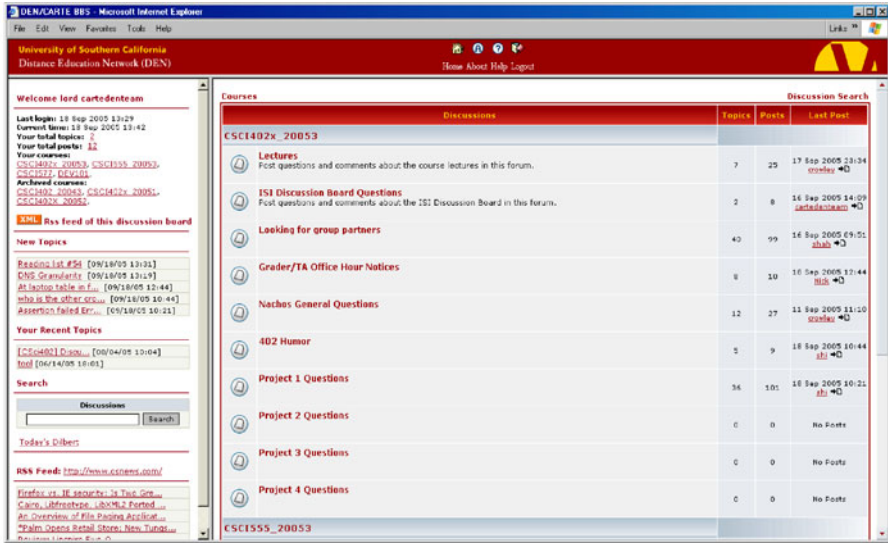
Finally we make use of all of these features as *predictive variables* and relate them to relevant class project grades. Our current results indicate that the degree of information provided to others, positive expressions, and earlier start on the project correlate with project performance.

The main contributions of this paper are: (1) presenting a framework for using online student activities as a predictor for project performance; (2) presenting approaches for capturing predictive variables that indicate student engagement, knowledge, and work pacing from online discussion data; and (3) evaluating the significance of the variables with a reasonably large dataset that discusses project related issues.

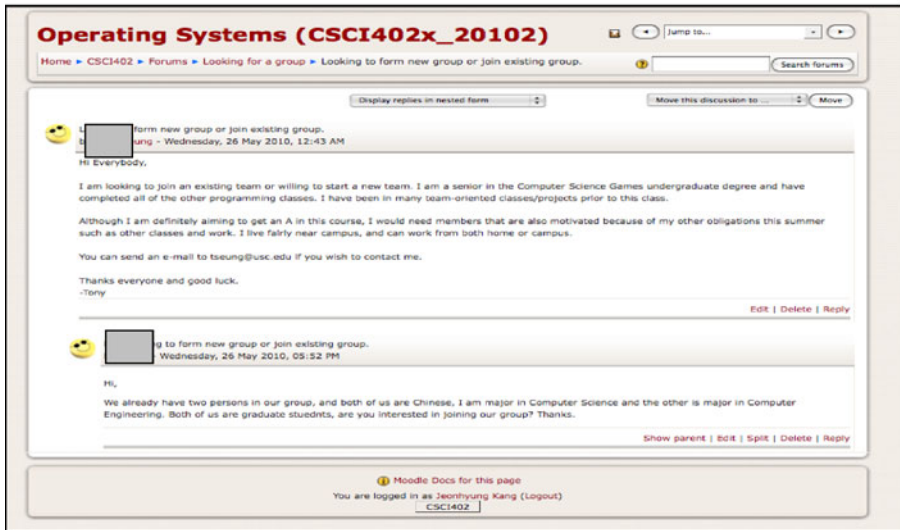
This paper is laid out as follows. First, we introduce online discussion data used for this study. Secondly, we discuss the nature of student online discussions and potential predictive variables. Thirdly, we present a framework for generating predictive variables that include data collection, data pre-processing, and data generation. Fourthly, we provide a detailed description of Sink/Source classifiers including data mining (classification and feature selection) needed for generating automated classifiers. Fifthly, we show the results from correlation and regression analysis, and identify significant variables in predicting student performance. We conclude the paper with a summary and directions for future research.

## Study Context: Online Discussions in Computer Programming Projects

Figure 1 shows example discussion boards from computer engineering courses. Instructors often use the board to promote discussions about projects or assignments. The online discussions or styles of interactions depend on how the class instructors set up the forums and how they use them for the class. As shown in Fig. 1(a), sub-forums can represent topics for individual projects or other issues relevant to the course such as questions on lectures. Each discussion consists of exchanges of messages, forming a discussion ‘thread’, as shown in Fig. 1(b). In Q&A style project forums, initial messages often present issues that the students encountered in performing the project. A message can receive more than one response, which can be posted by other students or the



(a) A phpBB discussion board for a class. Subforums can categorize discussion topics.



(b) Example discussion thread in Moodle

**Fig. 1** Example online discussion boards used by engineering courses. **a** A phpBB discussion board for a class. Subforums can categorize discussion topics. **b** Example discussion thread in Moodle

instructor. Long discussions can include clarifying questions, elaborating the initial problem, criticizing and suggesting answers, and acknowledging the answer (Kim et al. 2006). We are interested in understanding the patterns of student participation in online discussions, and how they can be used as a predictive model for related project performance. We expect that such predictive information can be used for alerting the instructor about students who may need more assistance or identifying strategies to improve the participation.

Our work makes use of data from undergraduate course discussion boards that are an integral component of an Operating Systems course in the Computer Sciences department at the University of Southern California. The course is held every semester and taught by the same instructor for the past 15 semesters. The instructor used two different platforms: phpBB and Moodle. We established a common discussion data representation and data transformation pipelines for handling discussion data from multiple different data sources. The students perform four group-based programming projects where the group size is no more than three. The group projects make up 40 % of the final grade. We studied data from recent eight semesters from the same course. Discussion is optional and not used for grading.

Among the 338 groups enrolled, 240 groups participated in discussions as shown in Table 1. We selected 173 groups who are active (posted more than three messages) for this study. In Table 1, ‘group’ participation means participation by at least one member of the group. About 71 % of the groups used the discussion board as means for exchanging information about their projects. In about 27 % of the groups, one or two students did not directly participate in online discussions and their ‘representatives’ interact with other teams. Students participated both in the discussions created by their own group and those generated by other groups. As our work focuses on group work performance, we treat each group as a unit. The instructor assigned the same grade to all members of each group.

## Capturing Predictive Variables from Online Discussion Data

This section describes a set of predictive variables that we can capture from student discussion data. In particular, we discuss a set of variables that indicate student engagement, knowledge and skills, and work patterns.

### Degrees of Participation

The simplest approach for assessing student activities from discussion participation is counting the number of messages. The degree of participation indicates their engagement

**Table 1** Forum participation of individual and group per semester ( $N = 240$ )

	Semester	Participation ratio	
		Individual	Group
	2006 Spring	0.62 (36/58)	0.83 (24/29)
	2006 Fall	0.55 (46/83)	0.78 (32/41)
	2007 Spring	0.40 (21/53)	0.67 (14/21)
	2007 Fall	0.65 (77/119)	0.78 (39/50)
	2008 Fall	0.50 (58/115)	0.73 (37/51)
	2009 Spring	0.48 (26/54)	0.80 (20/25)
	2010 Spring	0.49 (49/99)	0.73 (32/44)
	2010 Fall	0.41 (57/140)	0.55 (42/77)
	Total of participation ratio	0.51 (370/721)	0.71 (240/338)

*Participation ratio* # of students participated/ # of students enrolled in the discussion forum

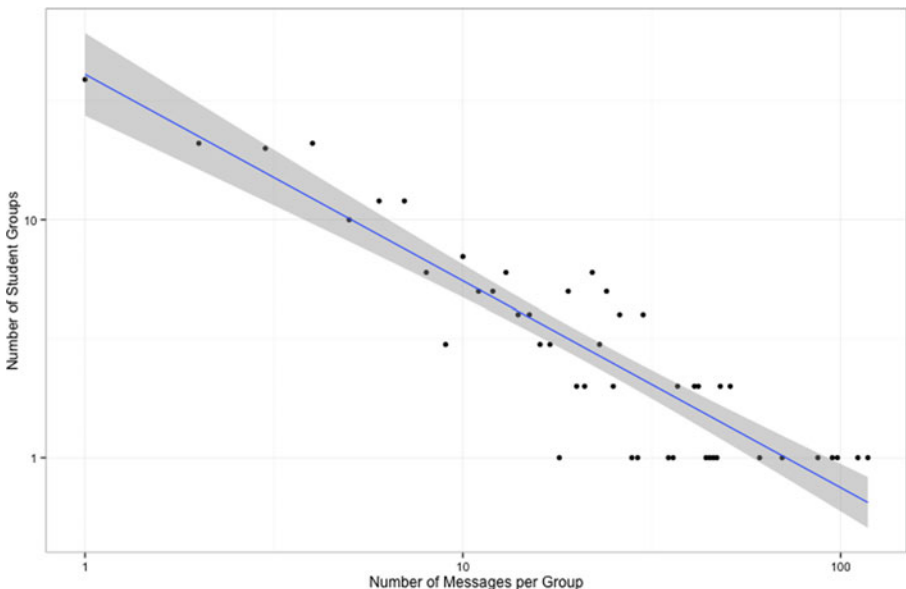
in information exchange. In quantifying the degree of participation, we can also consider the number of words, sentences, and paragraphs contributed by the student. Although simple, such features can provide hints on how actively the student or the group participates in the project.

Figure 2 shows the distribution of number of messages per student group ( $N=240$ ,  $\mu=13.17$ ,  $\sigma=18.70$ ). It follows a typical long-tail distribution in social media participation; many groups participate only a few times. However, 27 groups participated more than 30 times and produced 46 % of total messages (3,162 messages). We use the degree of participation as one of the predictive variables for related project performance.

In addition to the number of messages, we also use the number of initial messages and responses generated by the students. As initial messages often present problems or questions that arise during the project, the number of initial messages can show the degree of difficulty that the students have. However, it may also indicate leadership in starting discussions. The number of responses can, on the other hand, present the degree of assistance provided to others. However, since responses can present additional questions, more accurate classification of contribution type (information seeking vs. providing messages) is needed, as described below.

#### Contribution Type or Roles Played by the Student

Figure 4 shows an example discussion thread from our corpus. The discussion thread consists of a set of messages that are linked by a ‘reply-to’ relation. The discussion starts a problem description. In Message 1, student S1 presents specific issues in system calls. The message includes computer programming codes or error messages relevant to the project as well as English words. The original Message 1 contains error messages, as shown in Fig. 3. Our text pre-processing step replaced it with a token \$CODE\_BLOCKS\$.



**Fig. 2** Distribution of number of messages per student group (log-log scale)

```

/auto/home-scf-22/csci402/testgnu/decstation-ultrix/bin/cpp
-I./userprog -I./threads start.s
str.t.s/auto/home-scf-22/csci402/testgnu/decstation-ultrix/bin/as
-mips1 -o start.o str.t.sstart.s: Assembler messages:start.s:136:
Error: absolute expression required `addiu&apos;start.s:144:
Error: absolute expression required `addiu&apos;start.s:152:
Error: absolute expression required `addiu&apos;start.s:160:
Error: absolute expression required `addiu&apos;start.s:168:
Error: absolute expression required `addiu&apos;gmake:
*** [start.o]
Error 1

```

**Fig. 3** Example code block in student online discussions

The message poster, S1, plays a role of information seeker and the message can be classified as a Sink (information taking message). There can be more than one information seeker or provider in a thread. In this case, S2 and S3 try to help S1 by asking more information about the problem and providing suggestions. Their messages are classified as Source (information or suggestion generating a message) based on their roles within the discussion. However, within a short discussion thread, the role of a participant does not change very often. In our dataset, we observed that in 1,341 out of 1,345 threads, the participant roles remain the same within a thread.

Note that students' messages are often very informal and can contain spelling mistakes (e.g. "helpfule", "confussed"). There are many different ways that they express the same information. Also, information can be given in a form of a question rather than an answer, such as in message 3: "Do you have comments like//or/\* in start.s?". In order to generate meaningful features from the discussion data, proper data processing has to be performed as discussed in the next section.

### Linguistic Styles in Messages

In evaluating students' emotional expressions and semantic understanding, we made use of LIWC and Coh-Metrix measures that are relevant. As in the analysis of student written text, such as essay grading, the types of words the student use or the coherency of the text can be related to his/her writing skills (McNamara et al. 2010). Table 2 shows linguistic features of the messages shown in Figure 4 as examples. Higher scores on the

**Table 2** Linguistic features and roles of individual messages

Message ID	User ID	Information role		LIWC		Coh-Metrix		
		Sink	Source	Posemo	Negemo	LSA	Flesch	Cohesion
Message 1	S1	1	0	0.71	0	0.21	59.08	0.49
Message 2	S2	0	1	1.04	1.04	0.17	65.98	0.65
Message 3	S3	0	1	0	0	0.05	89.58	0.54
Message 4	S1	1	0	1.43	1.43	0.17	73.00	0.66
Message 5	S2	0	1	4.41	0	0.07	84.42	0.34
Message 6	S1	0	0	14.29	14.29	-0.02	100.00	0.25



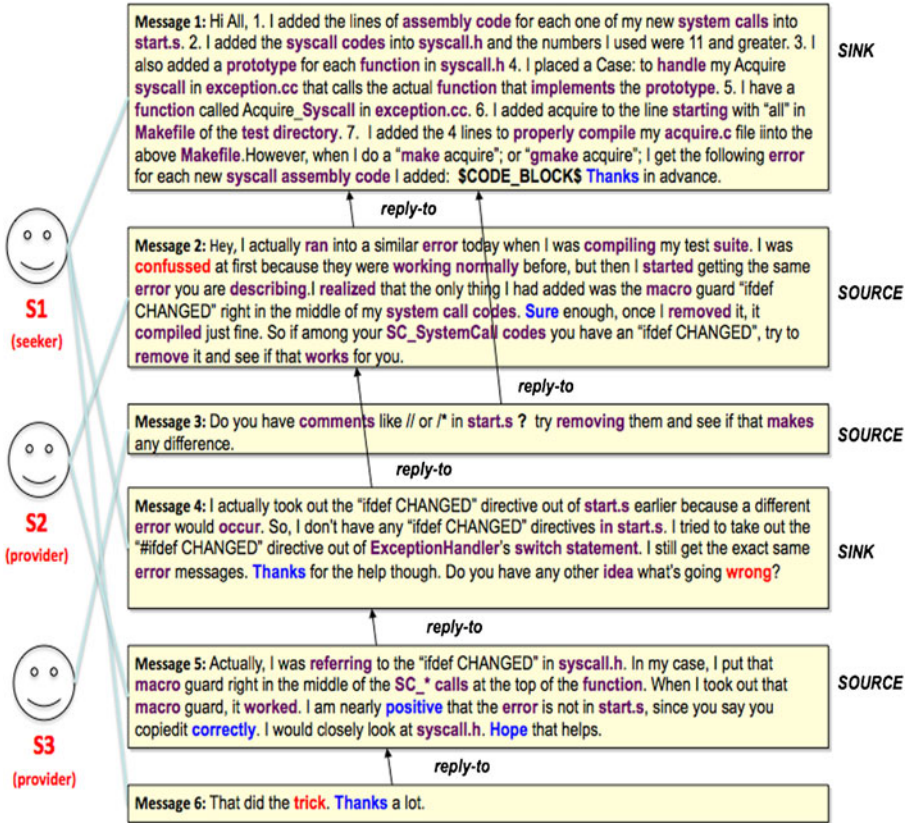


Fig. 4 Example discussion thread: blue, red, and violet color in a message represent positive, negative, and technical words respectively

Flesch reading test indicate that the text is easier to read. LSA (Latent Semantic Analysis) scores represent semantic coherence across the sentences. Messages 3, 5, and 6 are more readable than others and have lower coherence among the sentences. Overall, the Sink messages in this case have higher LSA scores than others; a seeker (S1) may have written Message 1 and 4 carefully to describe her/his problems or issues. There is a statistically positive correlation between Flesch scores and Cohesion ( $r = .64, n = 173, p = 0.01$ ). More cohesive text in discussions may be easier to read while cohesion gaps in text force the reader to generate inferences to bridge those gaps (McNamara and Graesser 2010; McNamara and Kintsch 1996). In addition, depending on the domain or the context of the discussion, the degree of technical term or domain term use may indicate how well the student knows related concepts.

Emotion Expressions

Positive expressions indicate student attitudes toward the subject that they discuss. Also polite expressions may show her/his relations with other discussants. For example, messages 1 and 5 contain more positive words than others ("thank", "correct"). Other linguistic features such as certainty and achievement may be important in characterizing

the student work and degree of confidence. We use the LIWC dictionary (Pennebaker et al. 2007) in classifying emotional and psychological words. Table 3 shows the distribution of frequent emotional and psychological LIWC words that appear in our dataset. Although, some of the terms such as “create” and “interrupt” (i.e. create/interrupt a thread) can be associated with our domain terms rather than emotional expressions, many LIWC terms still can be related to positive and negative emotions expressed by the discussants.

In Fig. 4, negative and positive emotion words in the messages are highlighted in red and blue respectively. The LIWC scores in Table 2 capture negative or positive words used in the discussion. However, the example also highlights challenges in identifying the actual emotion using LIWC; “trick” does not represent a negative emotion in this case.

### Participation Time and Procrastination

Figure 5 contrasts two students’ participation patterns. Each dotted vertical line represents a project deadline. Student A contributes to the discussion board throughout the four project periods. That is, the student asks questions or responds to other students from the beginning of the project period and maintains his/her activity throughout. However, some students participate only when the due date is close. For example, student B’s pattern represents a case where most of his/her posts were made close to the deadline. Such discussion participation patterns can expose the time when they worked for the project. That is, the more posts they make earlier, the earlier they may have started to work. The students who procrastinate may perform poorly for the projects.

We created a new measure for the distance between the student work time (or the posting time) and the deadline. The Average Posting Time To Deadline (APTTD)

**Table 3** Frequent emotional and psychological words

Positive emotion		Negative emotion		Certainty		Achievement	
Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
create	794	problem	1089	all	1510	able	2316
like	754	low	781	sure	558	king	1252
ease	727	numb	677	correct	494	create	794
thank	649	fault	581	must	332	first	479
sure	558	nag	437	never	249	acquire	405
please	494	interrupt	383	every	223	try	347
value	462	wrong	267	exact	180	work	344
fine	297	fail	262	always	165	fail	262
good	281	destroy	165	true	157	gain	256
well	265	argue	146	fact	150	require	248
ok	172	evil	109	certain	99	solution	132
free	166	sorry	82	clear	98	better	126
ready	159	worry	69	total	93	creative	117

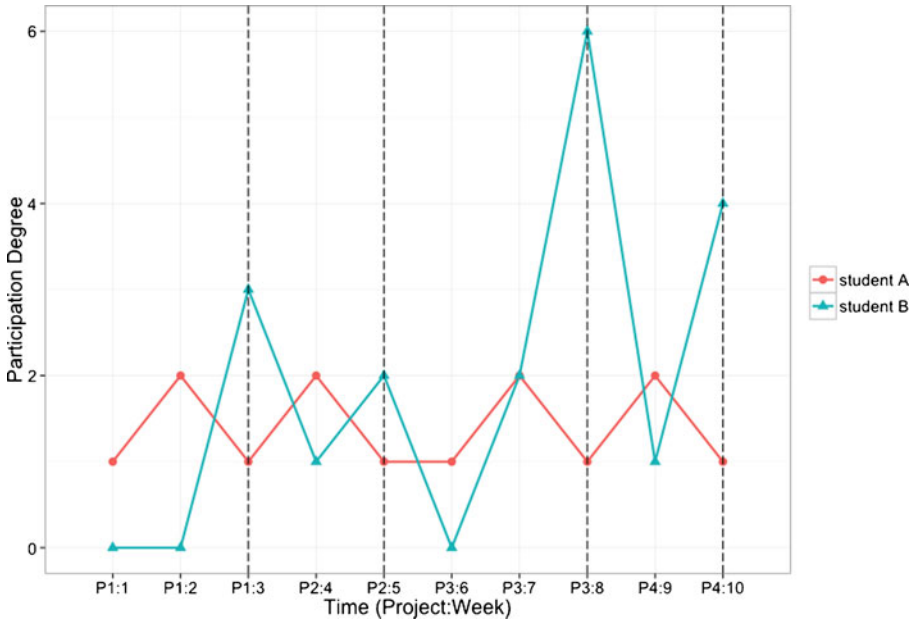


Fig. 5 Degree of participation over time

computes the average distance between the student postings and the deadlines. First, the Posting Time To Deadline (PTTD), when student  $s$  posts  $i$ -th message, is defined as follows:

$$PTTD_S^{(i)} = \frac{end_p - msg_s^{(i)}}{end_p - start_p} \tag{1}$$

where  $end_p$  is the deadline of project  $p$ ,  $start_p$  is the start date of project  $p$ , and  $msg_s^{(i)}$  is the time when student  $s$  posts  $i$ -th message. The numerator is the elapsed time between a message posting time and its project deadline and the denominator is a normalization factor that makes the posting time comparable across different projects. The smaller the PTTD of a message, the closer it is to the deadline. If a student has small PTTD values, his or her participations were closer to the deadline, which can indicate procrastination. The APTTD of a student  $s$  is calculated by averaging the PTTDs of the student over the semester. The APTTD of a group can be calculated as the sum of the APTTDs of the group members divided by the group size.

We compared two different groups who posted the same number (10) of messages in different time periods. Figure 6 provides multiple different views of their participation times. Figure 6(a) and (b) show that all the PTTDs of Group A are greater than 0.5. However, most PTTDs of Group B are smaller than 0.5. In order to compare posting times in the four different project periods, we normalized the time periods with a value range ( $0 \leq t \leq 1$ ). Figure 6(c) contrasts participation in the first and the latter half of each project. All of Group A’s participations (red bars) occurred in the first halves of the projects while most messages from Group B (green bars) were in the latter halves. Finally, Figure 6(d) presents each group’s APPTDs of individual messages in a sequence. The first data point represents the PTTD of the first message;

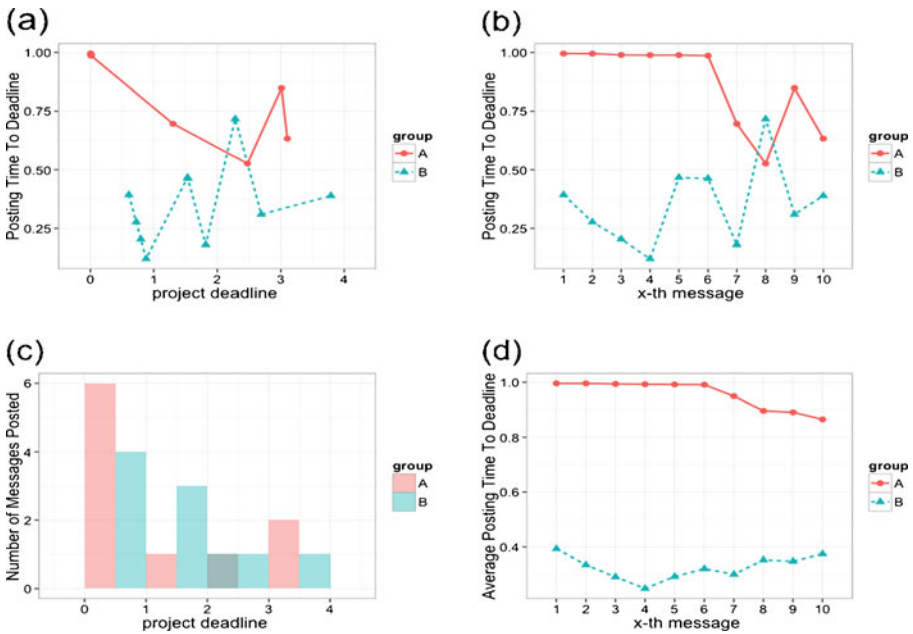


Fig. 6 Degree of participation over time

the second one is the APTTD of the first two messages, and so on. The final APTTD for Group A is 0.86 and for Group B is 0.37. Overall, Group A started to work on the project much earlier than Group B. In fact, we found that Group A received a higher grade than Group B (0.93 vs. 0.72).

### Procedures for Predictive Variable Generation

We introduce a procedural framework for generating each predictive variable that we described above and predicting student performance. Figure 7 illustrates the steps:

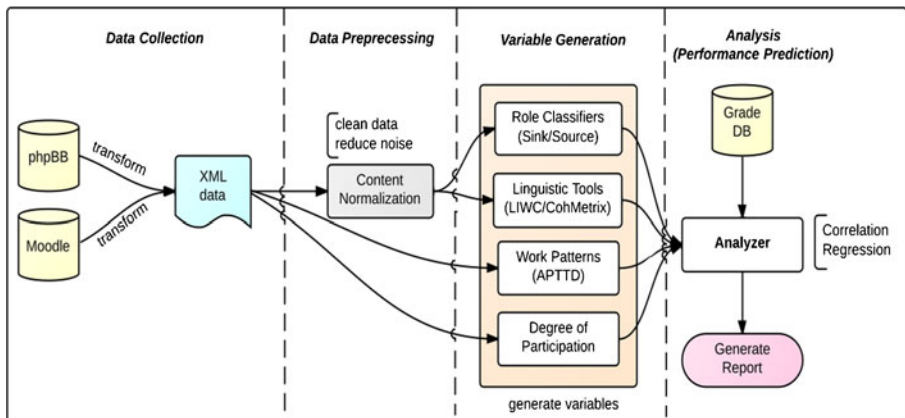


Fig. 7 A Procedural framework for predicting project performance using online discussion data

data collection, data pre-processing, predictive variable generation, and analysis (performance prediction).

### Data Collection

The class used phpBB (2006 ~ 2009) and then Moodle (2010 ~ current), which is an open source learning management system. The eight semesters' discussion data over 5 years have been collected from phpBB and Moodle databases. As shown in Fig. 7, we created a data processing pipeline for each platform so that they can be transformed into a consistent XML format that contains discussion posts, poster information, posting time, and so on.

### Data Pre-Processing

Student discussion data has a lot of noise including spelling mistakes, programming code, and quoted text. Also, students present similar information in many ways. Depending on the type of text analysis, in order to generate meaningful features, appropriate noise reduction and normalization have to be performed. For processing a large amount of discussion data, we implemented a fully-automatic data pre-processing procedure.

*Removing Unnecessary Content* There are two types: quotes (repetition of the replied-to message content inside the current post) and programming content (e.g., a block of C++) that are dominant in programming project discussions. For the LIWC and Coh-Metrix measures that rely on the number of words, we remove the copy of the replied-to message. Also, a large block of code that appears inside the text may confuse LIWC and Coh-Metrix, which expects normal English text as input. Although some editors allow users to distinguish quotes or programming codes by inserting special tags, not all users use such functions. In order to detect quotes of the replied-to messages, we used a text comparison tool called "google-diff-match-patch" (Google 2009). With this tool, we can identify the part that overlaps with the replied-to message. For detecting programming content, we developed a set of regular expressions to identify variable assignments, function definitions, function calls, etc. This approach eliminates programming content as shown in Fig. 3. The programming content in Message 1 in Fig. 4 is replaced with a code tag ("`$CODE_BLOCK`").

*Variance Reduction (Normalization)* The text written in discussion messages is often very informal and noisy with respect to grammar, syntax, and punctuation. In order to generate and run automated text classifiers, such as Sink and Source classifiers, our data pre-processing steps fix common spelling mistakes and abbreviations, convert contracted forms to their full forms, and transform informal words to formal words. For example, Fig. 8 shows that student written text may contain grammatical errors and noise (e.g. "shud" and "cant"). "cant" should be converted to "cannot" and "shud" should be corrected to "should." For another example, "I'm" is replaced by "I am", "you're" is replaced by "you are", and "we're" is modified to "we are". As other examples, "ya", "yea", and "yup" are all corrected to "yes." As identifying Sink

View previous topic :: View next topic	
Message	Author
<p>Posted: 13 Oct 2004 17:11</p> <p style="text-align: right;">nlp <input type="button" value="Reply"/> <input type="button" value="edit"/> <input type="button" value="X"/></p> <pre> /aut/home-scf-22/csci402/testgnu/decstation-ultrix/bin/ld -G 0 -c -I../userprog -I../threads start.o sleepingbarber.o -o sleepingbarber.coff /aut/home-scf-22/csci402/testgnu/decstation-ultrix/bin/ld: cannot open linker script file -I../userprog: No such file or directory gmake: *** [sleepingbarber] Error 1  i cant get my sleeping barbers test compiled???? what shud i do </pre> <p style="text-align: right;"><input type="button" value="email"/></p>	<p><b>Anonymous</b></p> <p>Joined: 28 Jan 2004 Posts: 135</p>

**Fig. 8** Example of student's noisy message text

and Source is agnostic on technical terms, in order to reduce the feature space, we replace them with a single token, \$tech\_term\$. More details on building Sink/Source classifiers are presented in next section.

### Predictive Variable Generation

Predictive Variable Generation in Fig. 7 mainly consists of four modules: Role (Sink/Source) classifiers, Linguistic Tools (LIWC and Coh-Metrix), Work Patterns, and Degree of Participation. Note that content in the message should be properly normalized before using Role classifiers and Linguistic Tools. From the above data pre-processing step, the input data is prepared for each module. The predictive variables are generated from four modules as follows: classified Role (Sink/Source), 80 LIWC variables, 60 Coh-Metrix variables, Work Patterns (APTTD), and Degree of Participation (i.e., the number of total, initial, and reply messages). Among the LIWC and Coh-Metrix metrics, we selected 20 LIWC and 10 Coh-Metrix metrics that are relevant to student project forums. For examples of LIWC, family, friends, and humans in social processes were excluded, as they are less relevant to technical discussions. Positive and negative emotional words that can capture affective tones in conversation were included. We also kept the words that represent cognitive processes such as causation, insight, and discrepancy. Among Coh-Metrix variables, readability, semantic indices, and situation model dimensions are chosen to capture contextual and semantic understanding. Technical domain terms are collected from the glossary in the textbook used by the class. The predictive variables used for our analysis are summarized below.

### Predicting Learner's Performance

After extracting selected predictive variables from the discussion data, correlation and regression analyses were performed with project grades.

**Table 4** Categories of variables

Categories	Metrics	Variables	Tool
Participation quantity	Quantitative	The number of Words, Sentences, Paragraphs, Messages (Total, Initials, and Replies)	Programming
Work pattern	Quantitative	APTTD (Average Posting Time to Deadline)	Programming
Technical	Quantitative	Technical terms	Dictionary
Information Roles	Qualitative	Sink, Source	Speech Act Classifiers
Linguistic	Quantitative	Past, Present, Future tense and Negations, Swear words	LIWC
	Qualitative	Flesch Reading Ease Score, Type-token ratio, Concreteness, Hypernym, Log frequency	Coh-Metrix
Emotional & psychological	Qualitative	Positive emotions, Negative emotions, Insight, Causation, Discrepancy, Certainty, Tentative, Inhibition, See, Time, Achievement, Assent	LIWC
Semantic factors	Qualitative	LSA sentence adjacent, LSA sentence all	Coh-Metrix
Situation model	Qualitative	Casual, Temporal, Spatial cohesion	Coh-Metrix

### Predictive Variables for Discussion Analysis

The categories of predictive variables used for the study are summarized in Table 4.

- *Participation quantity*: The total number of messages contributed, the number of Words, Sentences, Paragraph, and Messages.
- *Time*: APTTD (Average Posting Time To Deadline) is calculated by averaging PTTDs defined in Eq. (1).
- *Technical content*: Technical domain terms are captured from the domain dictionary or the glossary in the canonical document.
- *Participation types or information roles*: The number of initial messages (which start a discussion) and replies (which replies to other messages). The number of Sink (a message which requests information from others) and Source (a message which provides information to others).
- *Linguistic*: the percentage of tense (Past, Present, and Future verb), Negations, and Swear words, Flesch Reading Ease Score (ranging from 0 to 100), Type-token ratio (the number of unique words divided by the number of tokens in a message), Concreteness (mean of content words), Hypernym (mean hypernym values of nouns and verbs), Log frequency (log frequency of content words).
- *Emotional and Psychological*: the percentage of Positive and Negative emotions, Insight, Causation, Discrepancy, Certainty, Tentative, Inhibition, See, Time, Achievement, Assent.
- *Semantic factors*: LSA sentence adjacent (LSA cosines for sentence-to-sentence), LSA sentence all (mean LSA cosines for all sentence combinations, not just adjacent sentences).
- *Situation Model*: Casual (Ratio of causal particles to causal verbs), Temporal (Mean of tense and aspect repetition scores), and Spatial (Mean of location and motion ratio scores) cohesion.

For details on LIWC and Coh-Metrix variables, please see LIWC (Pennebaker et al. 2001) and Coh-Metrix (Graesser et al. 2004).

## Generating Machine Classifiers for Information Roles

For generating information roles, we have developed automatic machine learned classifiers. The classifiers are trained with human-annotated discussions from two semesters' discussion data. The 2006 spring and 2007 fall semester discussion data (898 messages) were randomly divided into two datasets: training dataset (628 messages) and test dataset (270 messages). The rest of the data (other semesters' data) is automatically labeled by the classifiers. This section describes the Sink/Source classifiers used for our analysis. In this section, the term 'feature' represents the features used for machine classifiers.

The discussion threads can be viewed as a special case of human conversation, and we adopted the theory of speech acts (Searle et al. 1980) to classify patterns of student's interaction. Identifying true information seeking (Sink) and providing (Source) dynamics in threaded Q&A is challenging: 1) noisy and informal discussion messages should be cleaned and normalized as mentioned before; 2) surface level features such question marks or interrogative words (who, what, where, when, how) are not enough to distinguish questions from answers and vice versa; 3) word  $n$ -gram approach considerably increases the dimensionality of the problem. The Sink/Source classifiers were built through three steps: feature generation, feature selection, and classification.

### Feature Generation

Before feature generation, we process the raw discussion data using the data preprocessing steps described above. We define two types of features; *message-level* features and *thread-level* features. The message-level features capture the standard  $n$ -grams in the message. We tried to reduce the variance of the features by categorizing "I" and "WE" into "IWE" or "HE" and "SHE" into "HESHE" and replacing "IS", "WAS", "ARE", "WERE" to "BE". Given a message, the content of the replied-to message provides useful hints about the role of the current message. For example, a response to a question (Sink) tends to be a Source. We include  $n$ -grams in the replied-to message as features of the current message. We use the following notation  $feature_{position}$  where the position represents either the replied-to (*rep*) or the current (*cur*) message, as shown in Table 5. As thread-level features, we also include author change information, the relative position of the message within the thread (e.g., first message, second message, etc.), and the features of the replied-to message. For example,  $Msg.Pos_{cur}$  and  $User.Pos_{cur}$  represent the position of the message and its poster (first participant, second participant, etc.) respectively. There were 30,044 features generated from the training dataset.

### Feature Selection

In order to improve the classification accuracy, we investigated ways to optimize features and eliminate irrelevant or redundant information using Information Gain (Dash and Liu 1997; Saeys et al. 2007; Yang and Pedersen 1997). The results from experiments for Sink/Source classifiers are shown in Fig. 9. To select the optimal number of features,



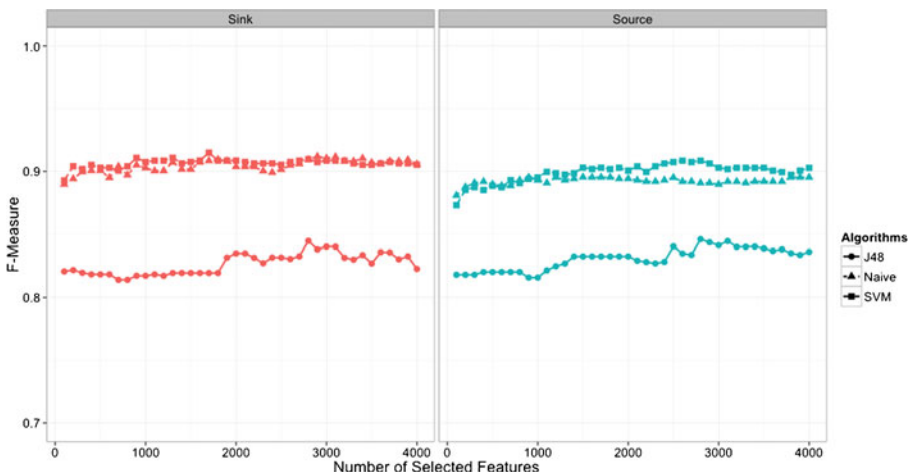
**Table 5** Top N-grams and structural information features sorted by information gain

Classifier	Unigram	Bigram	Trigram	Thread-level features
Sink	? <sub>cur</sub>	0.354 DO+IWE <sub>cur</sub>	0.069 DO+IWE+NEED <sub>cur</sub>	0.030 User.Pos <sub>cur</sub> 0.290
	? <sub>rep</sub>	0.242 IWE+BE <sub>cur</sub>	0.050 IWE+BE+GET <sub>cur</sub>	0.027 Msg.Pos <sub>cur</sub> 0.258
	IWE <sub>rep</sub>	0.173 CAN+IWE <sub>rep</sub>	0.047 WH+DO+IWE <sub>cur</sub>	0.022 User.Pos <sub>rep</sub> 0.140
	IWE <sub>cur</sub>	0.141 IWE+BE <sub>cur</sub>	0.045 DO+IWE+NEED <sub>rep</sub>	0.013 Msg.Pos <sub>rep</sub> 0.105
	BE <sub>rep</sub>	0.099 DO+IWE <sub>rep</sub>	0.038 IWE+BE+NOT <sub>cur</sub>	0.011
	YOU <sub>cur</sub>	0.095 YOU+CAN <sub>cur</sub>	0.034 YOU+CAN+NOT <sub>cur</sub>	0.010
	WH <sub>cur</sub>	0.073 BE+IT <sub>cur</sub>	0.033 YOU+NEED+TO <sub>cur</sub>	0.001
Source	? <sub>cur</sub>	0.290 DO+IWE <sub>cur</sub>	0.065 DO+IWE+NEED <sub>cur</sub>	0.028 User.Pos <sub>cur</sub> 0.303
	? <sub>rep</sub>	0.271 CAN+BE <sub>rep</sub>	0.051 WH+DO+IWE <sub>cur</sub>	0.021 Msg.Pos <sub>cur</sub> 0.251
	IWE <sub>rep</sub>	0.181 DO+IWE <sub>rep</sub>	0.047 IWE+BE+GET <sub>cur</sub>	0.019 User.Pos <sub>rep</sub> 0.160
	IWE <sub>cur</sub>	0.115 CAN+IWE <sub>cur</sub>	0.042 WH+DO+IWE <sub>rep</sub>	0.015 Msg.Pos <sub>rep</sub> 0.130
	YOU <sub>cur</sub>	0.102 IWE+BE <sub>cur</sub>	0.038 YOU+CAN+NOT <sub>rep</sub>	0.012
	BE <sub>rep</sub>	0.089 YOU+CAN <sub>cur</sub>	0.034 IWE+BE+NOT <sub>cur</sub>	0.011
	WH <sub>rep</sub>	0.082 BE+IT <sub>rep</sub>	0.029 YOU+NEED+NOT <sub>rep</sub>	0.011

we increase the number of selected features from 100 to 4,000 with an increment of 100. As a result, we use top 1,700 and 2,600 features for Sink and Source classifier, respectively. Some of the top N selected features are shown in Table 5. Most message-level features come from the current message. Some of the thread-level features were ranked high, including message/user positions in the thread.

Classification

A message can contain more than one role so we chose binary classification. We performed a 10-fold cross validation with the training dataset (628 messages) for selecting features and used the test dataset (270 messages) for predicting. All the



**Fig. 9** Feature selection using information gain for role classifiers

threads were annotated by hand beforehand: the Kappa scores of Sink and Source were 92.92 % and 95.95 % respectively. We then use J48 (open source Java implementation of C4.5), Naïve Bayes, and SVM (Support Vector Machine). Overall, SVM is better and less sensitive to the change of the number of selected features than J48 and Naïve Bayes. Therefore, we chose SVM for developing the final Sink/Source classifiers. We use F-measure (F-score), the harmonic mean of precision and recall, as a test's accuracy. The results of 10-fold cross validations reaches 0.929 for sink, 0.921 for source. The F-score of the role classifiers for sink and source on test dataset have improved from 0.871 to 0.914 and from 0.861 to 0.908 respectively by optimizing the features as summarized in Table 6.

## Results

This section presents correlation and regression analysis results.

### Correlation Analysis

Table 7 shows that 5 out of 46 independent variables are significantly related to the project grade (dependent variable). The correlation analysis revealed that Source (information providing role) is in fact a significant factor. Students who tend to answer others' questions may have understood the topic better, and produce better project outcome. Although we expected that low performing groups may ask more questions and seek more help than high performers, Sink does not have a significant coefficient. Surprisingly, the simple statistical information such as frequency of postings (Total and Reply) was related to the learner's performance. There has been controversy on whether such measures are effective (Palmer et al. 2007; Davies and Graff 2005; Bliss and Lawrence 2009).

Against our expectation that high performers tend to use more technical terms in their messages, the degree of technical terms was not significantly correlated with the performance. Among LIWC variables, only positive emotion was positively correlated with the project grade. Information providers may present positive attitudes and praise others more often than others. However, as positive words may be used by information seekers, as shown in M1 and M6 in Fig. 4, we plan to categorize different uses of positive words, such as praise vs. politeness.

**Table 6** The best F-measure on test dataset for role classifiers by algorithm

Algorithm	Sink classifier				Source classifier			
	Precision	Recall	F-score	Selected features	Precision	Recall	F-score	Selected features
J48	0.845	0.842	0.848	2800	0.845	0.849	0.847	2800
Naïve Bayes	0.908	0.914	0.911	2900	0.896	0.894	0.895	2700
SVM	0.910	0.918	<b>0.914</b>	1700	0.904	0.912	<b>0.908</b>	2600

**Table 7** Correlation between variables and grade among learners

Categories	Metrics	Variables	Tool	Correlation (Grade)			
Simple measures	Quantitative	Number of words	Programming	-.09			
		Number of sentences		-.10			
		Number of paragraph		-.08			
		Number of messages		.17*			
			Total	.13			
			Initial	.17*			
			Replies	.21**			
Time	Quantitative	APTTD	Programming	.08			
Technical	Quantitative	Technical terms	Dictionary	.03			
Speech act	Qualitative	Sink	Speech Act	.03			
		Source	Classifiers	.22**			
Linguistic	Qualitative	Past tense	LIWC	-.09			
		Present tense		.10			
		Future tense		.08			
		Negations		-.11			
		Swear words		-.09			
		Flesch Reading Ease Score		Coh-Metrix	-.08		
		Type-token ratio		-.06			
		Concreteness		-.08			
		Hypernym		-.08			
		Log frequency		-.07			
		Psychological		Qualitative	Positive emotions	LIWC	.16**
					Negative emotions		.10
					Insight		-.01
Causation	.05						
Discrepancy	.16						
Certainty	.10						
Tentative	.05						
Inhibition	-.02						
See	.06						
Time	.01						
Achievement	-.02						
Assent	.02						
Semantic	Qualitative		LSA sentence adjacent		Coh-Metrix		-.07
		LSA sentence all	-.05				
Situation model	Qualitative	Casual cohesion		.01			
		Temporal cohesion		-.06			
		Spatial cohesion		-.11			

$N = 173$ ; \* $p < .05$ ; \*\* $p < .01$

Linguistic variables (e.g., Type-token ratio and concreteness) in Coh-Metrix are not correlated with the project grade. As they are dependent on word counts, such

measures may not be effective for evaluating short messages. Contrary to our expectation, there is no correlation between the project grade and semantic (LSAs) or Situation model (Cohesions) from Coh-Metrix. We also did not observe correlation between the project grade and text coherence. Lastly, cohesion was also not a significant factor.

### Multiple Regression Analysis

In order to identify which variables explain the variance of our model, multiple regression analysis was conducted with the project grade as the dependent variable that is normalized by using a Z-score transformation. An analysis of variance test suggests that the regression model is significant,  $F(3, 169) = 17.08, p < 0.001$ , with 32 % variance in student group performance being explained by three predictors (The variables that were not significantly related to the grade were not included in the regression.) The multiple regression coefficient between the grade and the linear combination of the three predictors was  $R = .57$ .

The result of the multiple regression is summarized in Table 8. One pair of simple statistical counting such as Total and Reply was automatically dropped from the regression analysis because many sources are replies. Source has the largest regression coefficient,  $B = .47 (p < .001)$  and a significant correlation with the project grade. That is, the more information the group provides to other groups, the better grade they achieve. APTTD has the second largest regression coefficient,  $B = .20 (p < .001)$ . This is consistent with recent findings from other researchers (Michinov et al. 2011) that high procrastinators tend to get lower grades. Among the LIWC variables, the only significant variable was positive emotion. This suggests that LIWC that has been used mainly for behavioral or social psychology research may be not the best tool for analyzing our technical Q&A discussion data (Abe 2009). For using LIWC, additional filtering of domain terms may be needed as described earlier.

### Related Work

Prior studies to date have analyzed discussions in non-engineering courses or within integrated engineering courses (Lineweaver 2010; Baran and Correia 2009). We are interested in identifying indicators of student project performance within online discussions. Many studies of student discussions rely on labor-intensive manual discourse coding that necessarily limits sample size. For example, Perkins and Murphy (2006) were able to measure individual engagement and critical thinking

**Table 8** Summary of multiple regression analysis

Variables in the equation			
Variable	<i>B</i>	Std. error	Beta
Source	.47	.06	.48***
APTTD	.20	.07	.20**
Positive emotion	.02	.01	.13*

$R = .57; N = 173; *p < .05;$   
 $**p < .01; ***p < .001$

processes in online discussions by identifying student clarification, exploration-support-assessment, inference, and strategy. Similarly, Gunawardena, et al. (1997) examined learning and knowledge construction by identifying cognitive activities, arguments, resources, and changes in understanding. These are characteristics that we would like to be able to explore in later studies.

With respect to gender analysis, Jeong (2006) and Graddy (2006) found that gender differences in communication style did not produce significant differences in response patterns amongst students. Jeong's analysis was hand-coded based on a model of argumentation, while Graddy (2006) used isolated words and word group instances computed in Wordsmith. The data was limited to one or two courses, and neither study focused on achievement. For large-scale analyses, it is necessary to employ automated approaches such as the information role classification described above.

McLaren et al. (2007)'s work used machine learning techniques to evaluate discussions to provide 'Awareness Indicators' and facilitate moderation of discussions. They employed a similar annotation schema and built Q/A and critical reasoning classifiers. In contrast to the free-form course discussions comprising our corpus, their Digalo corpus was fairly well-structured with pre-defined problems. Text analysis work from Dönmez et al. (2005) and use of RapidMiner (formerly YALE) (RapidMiner 2007) for data mining are approaches that may complement our natural language processing (NLP) approaches.

In the field of dialogue analysis, there has been prior work on speech act framework and associated surface cue words (Carvalho and Cohen 2005; Samuel et al. 1998; Hirschberg and Litman 1993). Although the techniques are closely related to our Sink/Source analysis, they are evaluated on clean and coherent datasets. In contrast, student discourse is often syntactically noisy and grammatically incoherent and we have extended the data processing pipeline to accommodate them. Similarly, statistical discourse analysis techniques such as Latent Semantic Analysis and tools such as Coh-Metrix (Kolda and O'Leary 1998; McCarthy et al. 2006a, b), which identify cohesion, or the contextual usage of text, have been used to analyze writing styles, genre, and reading level. Our work builds on these results and we make use of metrics relevant to evaluating student project performance.

Peer interactions in collaborative dialogue or peer tutoring behavior have been analyzed to identify different linguistic features such as authority (Mayfield and Rosé 2011) or providing conceptual help (Walker et al. 2011). Participants' roles in collaborative games were analyzed using student chat data (Keshtkar et al. 2012). However such features were not fully used in evaluating student work or performance. We plan to explore opportunities to use these additional features such as evaluating leadership in performing the project.

In engineering courses, the focus has been on collaborative dialog analysis and computer supported collaborative learning (Cakir et al. 2005; Chi 2000; Scardamalia and Bereiter 1994; Stahl 2002; Soller and Lesgold 2003; Suthers and Hundhausen 2002; Kolodner and Nagel 1999; Hill et al. 2002). These results provide useful insights on student interactions and how technology can facilitate knowledge sharing and learning, but do not provide a broad analysis of discussion content for assessing factors that are relevant to project performance.

In supporting group programming projects in engineering courses, collaborative program design and implementation workspaces are increasingly explored and

analyzed (Bolanos and Sierra 2009). Our work provides additional measures that make use of online discussion participation information.

Instructional assessment tools have been developed to support formative assessment of student online discussion activities (Ma et al. 2010, 2011). We plan to explore opportunities to use our discussion analysis results to generate reports on students who need more assistant.

## Conclusion

We present a framework for using online student discussion participation as a predictor for class project performance. In order to assess student discussion participation, we evaluated dialogue and interaction features that have been useful for assessing student written text and collaborative dialogues. Building on prior work on assessing student written text and collaborative dialogue, we investigated indicators for student engagement, knowledge and work pacing and presented approaches for capturing such indicators. In particular, we present novel approaches for capturing student help seeking and information providing roles, and work pacing. In order to generate meaningful features from automatic text processing tools and classifiers, the raw discussion data was processed with various noise reduction and normalization steps. Automatic machine learned role classifiers were developed for efficient data processing. The final sink and source classifier accuracies reached 91.4 % and 90.8 % respectively. The resulting variables were used for a correlation analysis. Reducing noise and incoherency is one of the most challenging parts of the analysis, and we plan to continue to improve it.

The current results indicate that qualitative dialogue features such as the degree of information provided to others and how early students discuss their problems before the deadline are important factors in predicting the project grade. Other quantitative characteristics or linguistic variables do not seem to contribute much. The presented work focused on a computer science course and we plan to apply the same approach to discussions from other engineering courses.

We expect that this predictive information can be used for alerting the instructor about students who may need more assistance or identifying strategies to improve the participation. For example, based on our results, instructors can encourage students to help information seeking peers, or examine their work pacing more closely. We are planning to interview and study with the participating instructors.

We plan to perform more comprehensive analysis with diverse conversational or collaborative discussion features including number of conversation partners, degree of interactions with teachers, etc. as well as additional textual features described in the “Related Work”.

For future work, we are investigating several directions for extending and generalizing the presented assessment framework. First a more comprehensive set of student interaction information including social network features can be incorporated. For example, the degree of interaction with high performers and centrality in the communication network may be related to student performance. Second, the assessment can include other types of group activity information such as group meeting notes, project design document, etc. Such information can reveal effectiveness of

teamwork and quality of intermediate products. Finally, the procedural framework for analyzing online activity data can be extended for developing a formative assessment tool. Through dynamic and continuous analyses of student activity data, we can generate a “just-in-time” summary with which instructors can efficiently decide whom to help and what kind of help to provide.

**Acknowledgments** This work is supported by the National Science Foundation, CISE IIS (award #0917328) and REESE (award #1008747) programs.

## References

- Abe, J. A. A. (2009). Words that predict outstanding performance. *Journal of Research in Personality*, 43(3), 528–531.
- Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, 12(5), 385–390.
- Baran, E., & Correia, A.-P. (2009). Student-led facilitation strategies in online discussions. *Distance Education*, 30(3), 339–361.
- Bliss, C. A., & Lawrence, B. (2009). From posts to patterns: A metric to characterize discussion board activity in online courses. *Journal of Asynchronous Learning Networks*, 13(2), 15–32.
- Bolanos, D., & Sierra, A. (2009). Integrated software testing learning environment for training senior-level computer science students. *Software engineering: effective teaching and learning approaches and practices* (p. 233).
- Bransford, J. (2000). *How people learn: Brain, mind, experience, and school*. National Academies Press.
- Cakir, M., Xhafa, F., Zhou, N., Stahl, G. (2005). Thread-based analysis of patterns of collaborative interaction in chat. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AI-Ed 2005)* (pp. 120–127).
- Carvalho, V. R., & Cohen, W. W. (2005). On the collective classification of email speech acts. In *Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information retrieval* (pp. 345–352). ACM.
- Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology*, 5, 161–238.
- D’Mello, S., & Graesser, A. (2012). Language and discourse are powerful signals of student emotions during tutoring.
- Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1–4), 131–156.
- Davies, J., & Graff, M. (2005). Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4), 657–663.
- De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, 46(1), 6–28.
- Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In *Proceedings of 2005 Conference on Computer Support for Collaborative Learning* (pp. 125–134). International Society of the Learning Sciences.
- Erlin, Yusof, N., Rahman, A. A. (2009). Students’ interactions in online asynchronous discussion forum: A social network analysis. In *Proceedings of the 2009 International Conference on Education Technology and Computer, ICETC’09* (pp. 25–29). Washington, DC, USA. IEEE Computer Society.
- Felder, R. M., & Henriques, E. R. (1995). Learning and teaching styles in foreign and second language education. *Foreign Language Annals*, 28(1), 21–31.
- Ganapathy, C., Shaw, E., Kim, J. (2011). Assessing collaborative undergraduate student wikis and SVN with technology-based instrumentation: Relating participation patterns to learning. In *Proceedings of the American Society of Engineering Education Conference*.
- Google (2009). google-diff-match-patch. Online at <http://code.google.com/p/google-diff-match-patch/>.
- Graddy, D. B. (2006). Gender salience and the use of linguistic qualifiers and intensifiers in online course discussions. *The American Journal of Distance Education*, 20(4), 211–229.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2), 193–202.

- Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, 17(4), 397–431.
- Harman, K., & Koohang, A. (2005). Discussion board: A learning object. *Interdisciplinary Journal of E-Learning and Learning Objects*, 1(1), 67–77.
- Hill, A. W., Dong, A., Agogino, A. M. (2002). Towards computational tools for supporting the reflective team. In *Artificial intelligence in design'02* (pp. 305–325). Springer.
- Hirschberg, J., & Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3), 501–530.
- Jeong, A. (2006). Gender interaction patterns and gender participation in computer-supported collaborative argumentation. *The American Journal of Distance Education*, 20(4), 195–210.
- Kay, R. H. (2006). Developing a comprehensive metric for assessing discussion board effectiveness. *British Journal of Educational Technology*, 37(5), 761–783.
- Keshtkar, F., Morgan, B., & Graesser, A. (2012). Automated Detection of Mentors and Players in an Educational Game. In *EDM 212–213*.
- Kim, J., Chern, G., Feng, D., Shaw, E., Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*.
- Kolda, T. G., & O'leary, D. P. (1998). A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Transactions on Information Systems (TOIS)*, 16(4), 322–346.
- Kolodner, J. L., & Nagel, K. (1999). The design discussion area: A collaborative learning tool in support of learning from problem-solving and design activities. In *Proceedings of the 1999 Conference on Computer Support for Collaborative Learning* (p. 37). International Society of the Learning Sciences.
- Koschmann, T. E. (1996). *CSCL: Theory and practice of an emerging paradigm*. Lawrence Erlbaum Associates, Inc.
- Krupnick, C. G. (1985). Women and men in the classroom: Inequality and its remedies. *On Teaching and Learning*, 1(1), 18–25.
- Kumar, R., & Rose, C. P. (2011). Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies*, 4(1), 21–34.
- Lineweaver, T. T. (2010). Online discussion assignments improve students' class preparation. *Teaching of Psychology*, 37(3), 204–209.
- Ma, J., Shaw, E., Kim, J. (2010). Computational workflows for assessing student learning. *The 10th International Conference on Intelligent Tutoring Systems Conference*.
- Ma, J., Kang, J., Shaw, E., Kim, J. (2011). Workflow-based assessment of student online activities with topic and dialogue role classification. *Proceedings of the AI in Education Conference*.
- Mayfield, E., & Rosé, C. P. (2011). Recognizing authority in dialogue with an integer linear programming constrained model. In *Proceedings of Association for Computational Linguistics*.
- McCarthy, P. M., Graesser, A. C., McNamara, D. S. (2006a). Distinguishing genre using Coh-matrix indices of cohesion. In *16th Annual Meeting of the Society for Text and Discourse*. Minneapolis, MN.
- McCarthy, P. M., Lewis, G. A., Duffy, D. F., McNamara, D. S. (2006b). Analyzing writing styles with Coh-matrix. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 764–769).
- McLaren, B. M., Scheuer, O., De Laat, M., Hever, R., De Groot, R., & Rosé, C. P. (2007). Using machine learning techniques to analyze and support mediation of student e-discussions. *Frontiers in Artificial Intelligence and Applications*, 158, 331.
- McNamara, D. S., & Graesser, A. C. (2010). *Coh-matrix: An automated tool for theoretical and applied natural language processing. Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.
- McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3), 247–288.
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1), 57–86.
- Michinov, N., Brunot, S., Le Bohec, O., Juhel, J., & Delaval, M. (2011). Procrastination, participation, and performance in online learning environments. *Computers & Education*, 56(1), 243–252.
- Murray, T., Woolf, B., Xu, X., Shipe, S., Howard, S., Wing, L. (2012). Supporting social deliberative skills in online classroom dialogues: Preliminary results using automated text analysis. *Proceedings of the Intelligent Tutoring Systems Conference* (pp. 666–668).



- Newman, D., Johnson, C., Webb, B., & Cochrane, C. (1997). Evaluating the quality of learning in computer supported co-operative learning. *Journal of the American Society for Information Science*, 48(6), 484–495.
- Osborne, M. (2003). Increasing or widening participation in higher education?—a European overview. *European Journal of Education*, 38(1), 5–24.
- Palmer, S., Holt, D., & Bray, S. (2007). Does the discussion help? the impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology*, 39(5), 847–858.
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates.
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC*. LIWC. Net: Austin, TX.
- Perkins, C., & Murphy, E. (2006). Identifying and measuring individual engagement in critical thinking in online discussions: An exploratory case study. *Journal of Educational Technology and Society*, 9(1), 298.
- Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In Privacy, security, risk and trust (PASST), 2011 I.E. Third International conference on Social Computing (SocialCom) (pp. 180–185). IEEE.
- RapidMiner (2007). Rapidminer. Online at <http://rapid-i.com/>.
- Ravi, S., & Kim, J. (2007). Profiling student interactions in threaded discussions with speech act classifiers. *Frontiers in Artificial Intelligence and Applications*, 158, 357.
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.
- Samuel, K., Carberry, S., Vijay-Shanker, K. (1998). Dialogue act tagging with transformation-based learning. In *Proceedings of the 17th international conference on Computational linguistics* Vol 2 (pp. 1150–1156). Association for Computational Linguistics.
- Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *The Journal of the Learning Sciences*, 3(3), 265–283.
- Searle, J. R., Kiefer, F., Bierwisch, M. (1980). Speech act theory and pragmatics, vol 10. D.
- Soller, A., & Lesgold, A. (2003). A computational approach to analyzing online knowledge sharing interaction. In *Proceedings of artificial intelligence in education* (pp. 253–260).
- Stahl, G. (2002). The complexity of a collaborative interaction. *Proceedings of ICLS*.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist; American Psychologist*, 52(6), 613.
- Suthers, D. D., & Hundhausen, C. D. (2002). The effects of representation on students' elaborations in collaborative inquiry. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community* (pp. 472–480). International Society of the Learning Sciences.
- Walker, E., Rummel, N., Koedinger, K. (2011). Using automated dialog analysis to assess peer tutoring and trigger effective support. In *Artificial intelligence in education* (pp. 385–393). Springer.
- Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Machine Learning-International Conference* (pp. 412–420). Morgan Kaufmann Publishers, Inc.