Inferring Students' Sense of Community from Their Communication Behavior in Online Courses

Wen Wu
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
cswenwu@comp.hkbu.edu.hk

Li Chen
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Qingchang Yang
Adaptive Learning Center
Hong Kong Baptist University
Hong Kong, China
yqc@teamcrushsport.com

ABSTRACT

Sense of community is regarded as the reflection of students' feelings of connectedness with community members and commonality of learning expectations and goals. In online courses, sense of community has been proven to influence students' learning engagement and academic performance. Low sense of community is also one of the reasons for drop out. However, existing studies mainly acquire students' sense of community via questionnaires, which demand user efforts and have difficulty in obtaining real-time feeling during students' learning process. In addition, although communication is helpful to enhance students' sense of community, little work has empirically compared the impact of different online communication tools. In this paper, we are motivated to derive students' sense of community from their communication behavior in online courses. Concretely, we first identify a set of features that are significantly correlated with students' sense of community, which not only include their activities carried out in both synchronous and asynchronous online learning environment, but also their linguistic content in conversational texts. We then develop inference model to unify these features for determining students' sense of community, and find that LASSO performs the best in terms of inference accuracy.

CCS CONCEPTS

•Human-centered computing → User studies; •Applied computing → Collaborative learning;

KEYWORDS

Online learning; sense of community; prediction; synchronous/ asynchronous communication

1 INTRODUCTION

In recent years, online learning, which is defined as the process of using Internet to acquire knowledge, access learning materials, and interact with others [1], has become popular. According to the data collected by Class Central, by 2016, around 58 million students worldwide have taken at least one course, and the total number of courses has grown to 6,850.

In learning environment, sense of community is one of the popularly used metrics to measure students' feeling that they connect to community members in a course-based context and the feeling that the community helps them to acquire knowledge and meet learning goals [18]. In physical classroom, sense of community is shown to be related to students' learning perception and actual performance [9, 30]. In online courses, it also plays an important role. Specifically, students with high sense of community tend to be active in online learning, feel satisfied with academic programs, become interested in the studied course, and be motivated to accumulate course knowledge [11, 19, 34, 35], whereas those with low sense of community easily feel anxious or isolated [30], which is one of the reasons for drop out [20]. Therefore, knowing students' sense of community has the potential to provide them with more personalized learning supports so as to improve their learning effectiveness and potentially alleviate the high dropout issue.

However, the issue of how to obtain students' sense of community in online courses has not been well solved. The existing studies mainly rely on questionnaires to explicitly acquire students' sense of community (such as 20-item Rovai's Classroom Community Scale [18]), which not only demand high user efforts, but also have difficulty in obtaining students' real-time sense of community during their learning process. Another limitation is that although communication has been proven to be effective in improving students' social interaction and their feeling of connectedness in online courses [5, 8, 25], little work has empirically compared students' usage of different communication tools (such as chat room, discussion forum, note-taking facility) and explored what communication tools may be more helpful to enhance students' sense of community. Therefore, we have aimed to answer the following two research questions:

RQ1: *How do various communication tools affect students' sense of community in online courses?*

RQ2: *To what extent can sense of community be inferred from students' online communication behavior?*

In order to address the two questions, we conduct our experiment with 489 college students in an online learning system called "eBanshu" (www.ebanshu.com), which provides both synchronous and asynchronous online communication

UMAP '17, July 9-12, 2017, Bratislava, Slovakia
© 2017 ACM. 978-1-4503-4635-1/17/07...$15.00
DOI: http://dx.doi.org/10.1145/3079628.3079678

tools. Concretely, the synchronous communication tools include *chat room* for students to exchange messages in real time, *hand-up facility* for students to ask questions to instructors in the online class, *note-taking* and *note-sharing facility* that allow students to take notes and share their written notes with others. The asynchronous tools include *discussion forum* where students can ask and/or answer other students' questions, *material-sharing facility* for students to share their learning materials with others, and *assignment submission facility* for students to submit assignments to instructors.

We have first performed a correlation analysis to study the relationship between students' behavior in using communication tools and their sense of community. Particularly, the behavioral features not only include students' in-class and after-class activities (such as the numbers of messages posted in chat room and discussion forum respectively), but also qualitative linguistic characteristics embedded in textual contents. Based on the results, a total of 15 features are found significantly correlated with students' sense of community. Among them, there are 6 activity features (e.g., number of using hand-up facility), 5 content features (e.g., number of social process words in each chat message), 2 personal properties (i.e., pre-course interest and pre-course knowledge) and 2 environmental features (i.e., numbers of instructors' and classmates' activities in each lesson). We have then built inference model based on these features to predict students' sense of community. We have concretely tested six popular regression models and found LASSO shows the best accuracy.

In the following, we first introduce related work (Section 2). We then give the details of our experimental setup (Section 3) and results analysis (Section 4). We finally discuss the implications and draw a conclusion (Section 5 and 6).

2 RELATED WORK

2.1 Sense of Community

The most widely accepted definition of sense of community was proposed by McMillan and Chavis [10] in 1986 based on [21, 28]. Their definition is "a feeling that members have of belonging, a feeling that members matter to one another and to the group, and a shared faith that members' needs will be met through their commitment to be together".

In the education domain, Rovai [18] defined classroom community (i.e., a community of learners) as "a feeling that members have of belonging, a feeling that members matter to one another and to the group, that they have duties and obligations to peers and to the school, and that they possess shared expectations that members' educational needs will be met through their commitment to shared goals". It consists of two components: *Connectedness* - feeling of connectedness with community members; *Learning* - commonality of learning expectations and goals. Specifically, *Connectedness* is the feeling of belonging and acceptance of bonding relationships. *Learning* is the feeling that knowledge and meaning are actively constructed within the community, that the community enhances the acquisition of knowledge and understanding, and that members' learning needs are satisfied.

2.2 Sense of Community and Learning

The sense of community within the physical classroom was shown important, because it is significantly correlated with students' classroom attitudes, perception of learning, and actual academic performance [9]. If students fail to feel the sense of community, they are likely to be anxious, defensive and not willing to take the risks involved in learning [30].

Recently, the importance of sense of community on online learning has been explored. It was shown in [19, 27, 34, 35] that online learners who have stronger sense of community tend to perceive greater cognitive learning, have higher satisfaction with their academic programs, accumulate more knowledge and achieve better academic performance. It was also found that sense of community could affect the retention rate, which may alleviate the dropout issue [11, 32, 35, 37].

In terms of how to build sense of community in the online environment, computer-mediated communication tools have been shown useful [5, 8, 25]. For instance, Sveningsson [25] observed that students' sense of community is closely related to their usage of web chat, whereas Dawson [5] found that the online forum discussion could facilitate the development of a strong community. However, the limitation is that little work has empirically compared different communication tools in terms of their effect on enhancing students' sense of community in online courses.

2.3 Measurement of Sense of Community

Rovai's Classroom Community Scale [18] is used to measure students' sense of community in virtual classrooms. It consists of 20 statements (each statement is rated on a 5-point Likert scale from 1 "strongly disagree" to 5 "strongly agree"), among which 10 statements are related to *Connectedness* (e.g., *"I feel connected to others in this course"*), and 10 are related to *Learning* (e.g., *"I think that this course results in only modest learning"*).

However, using questionnaires to acquire students' sense of community unavoidably demands user efforts. In order to solve the problem, Shea [23] used students' demographic information and teaching presence (i.e., process of design, facilitation, and direction of cognitive and social processes) to predict their sense of community. But the limitation of his method is that the sense of community cannot be obtained in real time during their learning process. In addition, little work has in depth studied the role of students' online communication behavior in predicting their sense of community.

We are thus interested in not only exploring what communication tools can be more helpful to enhance students' sense of community, but also investigating how to infer students' real-time sense of community from their communication behavior in online courses.

3 EXPERIMENTAL SETUP

3.1 Materials and Participants

In order to answer our research questions (see Section 1), we conducted an experiment on eBanshu online learning system, which was released in 2013 and has been used by more than

20 universities in China with over 33,000 students who have enrolled on 100 courses so far. In this website, instructors can use video cameras and digitizers (for writing notes) to give real-time lectures. In the online class, students can communicate with instructors and peers through a text chat room, ask or answer questions by using the hand-up facility, and take notes and share them (see Figure 1). After class, they can leave messages in a course-based discussion forum, share learning materials, and submit assignments. These communication tools are provided for students to freely use, not counted in their final assessment.

From March to June 2015, a total of 1,559 students, from Hebei Normal University in China enrolled in 16 elective courses in 3 different subject types: liberal arts (including 9 courses, e.g., "Comparative Literature"), science (6 courses, e.g., "Discrete Mathematics"), and engineering (1 course, "Microcomputer Principles and Interface Description"). Each student enrolled in one course, and the average number of enrollments per course is 97.3 (min=50, max=209, st.d.=42.2). Each course lasted for 12 weeks, with 2 lessons given every week (each lesson took 1 hour). At the end, students received credit if they passed the assignments and examinations. We sent survey invitation to all of the 1,559 students before they attended class, of whom 508 students accepted. After filtering out incomplete and invalid answers that they gave to the survey questions, we finally got 489 students' data (with 408 females). Their ages range from 20 to 25 and the students are from 11 different majors (e.g., English, Physics, Mathematics, Pedagogy).

3.2 Procedure and Measurement

3.2.1 User survey. In the questionnaire asked before they took course, we included some questions about the student's personal properties, such as **age**, **gender**, **pre-course interest** (*"Before learning, my interest in the course is (): from 1 'very low' to 5 'very high'"*) and **pre-course knowledge** (*"Before learning, I have obtained () of the needed knowledge: from 1 'none' to 5 'all'"*). Besides, some course-related factors are also included [15, 17]: **subject type of each course** (i.e., liberal art, science, or engineering), **course structure** (the number of sections in each course), and **assessment structure** (the number of assignments).

When students finished the course, we asked them to fill in a post-course questionnaire in order to acquire their **sense of community**. We assessed the student's sense of community with Rovai's 20-statement Classroom Community Scale (as introduced in Section 2.3), which reaches satisfactory convergent and discriminant validity [18]. In addition to its original classification (*Connectedness* and *Learning* two sub-scales [18]), we proposed a new classification in order to assess students' perception of *Interaction with Instructor (InterInstructor, for short)* and *Interaction with Other Students (InterStudent)*. Specifically, *InterInstructor* is assessed by 10 items selected from the Classroom Community Scale (e.g., *"I feel that I am encouraged to ask questions"*) and 1 new item (*"I feel I can actively interact with my instructor during the online course"*); and *InterStudent* is also assessed by 11 statements, including the other 10 items in the Classroom Community Scale (e.g., *"I feel that students in this course care about each other"*) and 1 new item (*"I feel I can actively interact with other students during the online course"*).

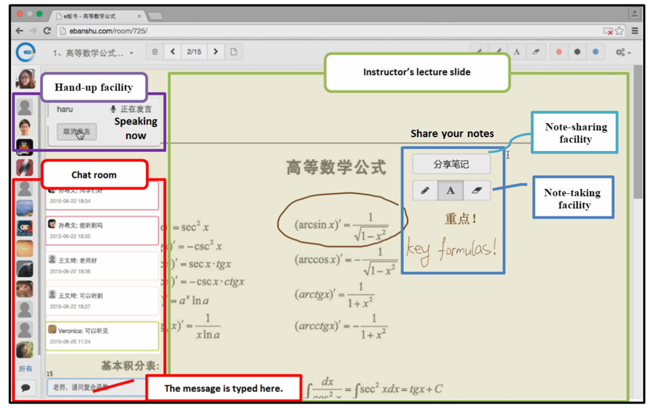

Figure 1: Snapshot of the synchronous instruction interface in eBanshu (www.ebanshu.com).

Table 1: List of students' online behavioral features

Activity features	**In-class activities**#	Frequency of using hand-up facility
		Number of messages posted in chat room
		Frequency of taking notes
		Frequency of sharing notes
		Class attendance rate
	After-class activities	Frequency of sharing learning materials
		Number of messages posted in discussion forum
		Assignment submission rate
Content features+	**Message length**	Number of words per sentence
	Psychological presence	Number of social process words per message
		Number of affective process words per message
		Number of cognitive process words per message
	Task engagement	Number of fully-engaged sentences
		Number of somewhat-engaged sentences
		Number of disengaged sentences

Each in-class activity (except class attendance rate) is measured in terms of both average number per lesson and total number during the whole course.
+ Content features are extracted from students' messages posted in chat room and discussion forum.

3.2.2 Objective online learning behavior. eBanshu system automatically recorded students' learning behavior in a log file, which includes not only the activities they carried out in class and after class, but also their text messages posted in chat room and discussion forum.

From the log file, we extracted two types of features: activity features and content features (see Table 1). The activity features are further divided into two categories: **in-class activity features** that include students' attendance rates, frequency of using hand-up facility, number of messages posted in chat room, number of notes taken in class, and frequency of sharing notes; **after-class activity features** that include students' assignment submission rates, frequency of sharing learning materials, and number of messages posted in discussion forum.

The content features include **sentence length**, **psychological presence**, and **task engagement** of messages students posted in chat room and discussion forum. To be

Table 2: Measurement of message content's psychological presence and task engagement level (the coding process for psychological presence is referred to [14])

Psychological presence	
	Coding
Social presence	Occurrence of social process words in CLIWC[#] sub-categories: *social* (e.g., "talk"), *friend* (e.g., "buddy"), *family* (e.g., "daughter"), and *human* (e.g., "adult").
	Occurrence of affective process words in CLIWC sub-categories: *affect* (e.g., "happy"), *positive emotion* (e.g., "nice"), *negative emotion* (e.g., "hurt"), *anger* (e.g., "hate"), *anxiety* (e.g., "worried"), and *sadness* (e.g., "sad").
Cognitive presence	Occurrence of cognitive process words in CLIWC sub-categories: *insight* (e.g., "think"), *causation* (e.g., "because"), *discrepancy* (e.g. "should"), *tentative* (e.g., "perhaps"), *certainty* (e.g., "always"), *inhibition* (e.g., "constrain"), and *inclusive* (e.g., "include").
Task engagement	
Fully-engaged	Occurrence of words/phrases that are closely related to learning (e.g., "assignment", "exam").
Somewhat-engaged	Occurrence of words/phrases that are somewhat related to learning (e.g., "ask for leave", "technical support").
Disengaged	Occurrence of words/phrases that are unrelated to learning (e.g., greeting words like "hello", modal particle words like "wow").

[#] CLIWC is short for "Chinese Linguistic Inquiry and Word Count" dictionary.

Table 3: Students' sense of community, personal properties, and course-related features

Sense of community	Overall	Mean=3.51 (st.d.=0.56)
	Connectedness	Mean=3.59 (st.d.=0.52)
	Learning	Mean=3.43 (st.d.=0.71)
	InterInstructor	Mean=3.52 (st.d.=0.66)
	InterStudent	Mean=3.50 (st.d.=0.54)
Personal properties	Age	20-25
	Gender	Female: 409 (83.6%); Male: 80 (17.4%)
	Pre-course interest	Mean=3.77 (st.d.=0.79)
	Pre-course knowledge	Mean=2.44 (st.d.=0.69)
Course-related features	Course structure	Mean=17.17 (st.d.=15.44)
	Assessment structure	Mean=6.00 (st.d.=9.77)

specific, sentence length is taken as a manifest indicator of students' sustained interaction [22]. Psychological presence evaluates whether students' online communication content can foster collaborative and meaningful learning, which is defined in two categories [13]: *social presence* (the degree of awareness of others in an interaction) and *cognitive presence* (the extent of both reflection and discourse in the construction of meaningful outputs). To encode the psychological presence of each message, we adopted a popularly used text analysis tool, Chinese Linguistic Inquiry and Word Count (CLIWC) dictionary [14]. If a word (in a message) belongs to "social process" or "affective process" (see the coding in Table 2), it is taken as the indicator of *social presence* [13]. Otherwise, if the word is coded as "cognitive process", it is classified as *cognitive presence* [13]. As for task engagement, it measures whether the posted message is related to the course content [4]. Each message sentence's task engagement level was manually determined by counting the occurrences of learning-related word/phrase. If the sentence contains word/phrase like "assignment" or "exam", it is classified as "fully-engaged". If it contains word/phrase such as "ask for leave" or "technical support", it is classified as "somewhat-engaged". Otherwise, if the sentence contains word/phrase that is not relevant to the learning task (such as "hello" or "wow"), it is classified as "disengaged". The definition of each engagement level is given in Table 2.

Table 4: The results of Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) test on sense of community overall scale and sub-scales

	K-S			S-W		
	Stat.	df	sig.	Stat.	df	sig.
SC_Overall	0.071	489	0.000	0.990	489	0.002
SC_Connectedness	0.089	489	0.000	0.968	489	0.000
SC_Learning	0.110	489	0.000	0.956	489	0.000
SC_InterInstructor	0.097	489	0.000	0.983	489	0.000
SC_InterStudent	0.102	489	0.000	0.964	489	0.000

4 RESULTS AND ANALYSIS

4.1 Data Overview

We are interested in verifying whether students' behavior in using different communication tools is correlated with their sense of community. Before reporting the correlation results, we first describe our collected data. In terms of our participants' sense of community (see Table 3), the reliability test of our used Rovai's Classroom Community Scale shows that its internal consistency coefficient (Cronbach's alpha) is 0.888, and the coefficients of the sub-scales *Connectedness*, *Learning*, *InterInstructor*, and *InterStudent* are 0.724, 0.811, 0.747, and 0.803 respectively. These values are all above 0.70, suggesting that the corresponding statements have satisfactory internal validity [12]. From Table 3, we see that the mean value of the *Overall* scale of sense of community (with 20 statements) is 3.51 out of 5 (st.d.=0.56). Regarding the sub-scales, the mean values of *Connectedness* (with 10 statements), *Learning* (with 10 statements), *InterInstructor* (with 11 statements), and *InterStudent* (with 11 statements) are 3.59 (st.d.=0.52), 3.43 (st.d.=0.71), 3.52 (st.d.=0.66), and 3.50 (st.d.=0.54) respectively. Besides, in order to choose the appropriate correlation measurement, we check the normality of sense of community scales using the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) tests. According to the results (see Table 4), both *Overall* scale and four sub-scales are not normally distributed at the 0.05 significant level ($Sig.<0.05$). In addition, it can be seen in Table 3 that most of students were interested in the enrolled courses (pre-course interest: mean=3.77, st.d.=0.79) and had few prior course knowledge (mean=2.44, st.d.= 0.69). As for course structure and assessment structure, they are respectively 17.17 (the average number of sections across all courses) and 6 (the average number of assignments).

The results of analyzing students' activities and message contents are given in Table 5. The average course attendance rate is 99.2%, indicating that the students took the majority of online lessons. Moreover, during the whole course, 98.8% of students posted at least one message in chat room and 59.7% had experience of using the hand-up facility. Additionally, although over half of students (57.1%) took at least one note in the online class, only 18.0% shared their written notes with others for at least one time. After class, the average assignment submission rate is 78.7%. Besides, 77.3% of the students shared their learning materials at least once. Relatively, the percentage of students who used discussion forum is lower, with 37.7%. On the other hand, the average numbers of activities among all students in using these

Table 5: Statistical results of analyzing students' activities and message contents

Behavioral features	Results	
Activity features	# (%) of st -udents who carried out the activity	# of activi -ties per stu- dent during whole course
# of using hand-up facility	292 (59.7%)	Mean=1.4
# of chat messages	483 (98.8%)	Mean=57.3
# of taking notes	279 (57.1%)	Mean=7.3
# of sharing notes	88 (18.0%)	Mean=0.97
Average course attendance rate: 99.2%		
# of material sharing	378 (77.3%)	Mean=2.8
# of forum messages	184 (37.7%)	Mean=0.7
Average assignment submission rate: 78.7%		
Content features	**Chat room**	**discussion forum**
# of words per sentence	Mean=3.60	Mean=4.41
# of social process words per message	Mean=0.46	Mean=1.47
# of affective process words per message	Mean=0.19	Mean=1.14
# of cognitive process words per message	Mean=0.81	Mean=4.47
% of fully-engaged sentences	16.5%	46.3%
% of somewhat-engaged sentences	23.5%	28.0%
% of disengaged sentences	60.0%	25.7%

Table 6: The results of Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) test on students' online activities

	K-S			S-W		
	Stat.	df	sig.	Stat.	df	sig.
# of using hand-up facility	0.307	489	0.000	0.589	489	0.000
# of chat messages	0.213	489	0.000	0.660	489	0.000
# of taking notes	0.384	489	0.000	0.284	489	0.000
# of sharing notes	0.415	489	0.000	0.249	489	0.000
Average course attendance rate	0.134	489	0.000	0.907	489	0.000
# of material sharing	0.185	489	0.000	0.830	489	0.000
# of forum messages	0.325	489	0.000	0.571	489	0.000
Average assignment submission rate	0.233	489	0.000	0.750	489	0.000

tools show that the frequency of posting messages in chat room is largely higher than those of others, with mean 57.3, vs. average 7.3 times of taking notes, 2.8 times of sharing materials, 1.4 times of using hand-up facility, 0.97 times of sharing notes, and 0.7 messages posted in discussion forum.

Therefore, the above results demonstrate that in online class, our studied students were more active in communicating with others synchronously through chat room, followed by asking question to instructors directly through the hand-up facility. After class, they more frequently shared learning materials, but the frequency of posting content in discussion forum was relatively low. Table 6 shows the results of K-S and S-W test, indicating that all of the activity features are not normally distributed ($Sig.$<0.05).

As for message content (see Table 5), the average sentence length of messages in chat room is significantly shorter than that in discussion forum (3.60 vs. 4.41, $p < 0.05$ via two-tailed paired t-test), which indicates that students like to write shorter sentence during synchronous communication. In addition, although the quantity of messages posted in discussion forum is lower than that in chat room, the quality seems better. Specifically, in discussion forum, students used more social process words (mean=1.47 vs. 0.46 in chat room messages, $p > 0.05$), affective process words (1.14 vs. 0.19, $p > 0.05$), and cognitive process words (4.47 vs. 0.81, $p < 0.05$), which indicates they prefer to show their psychological presence using the asynchronous communication tool. Another phenomenon is that both types of messages (in discussion forum and chat room) include more cognitive process words than social and affective process words, implying that through sustained communication, students are more inclined to construct meaning (i.e., exerting cognitive presence) than to enhance interaction (i.e., exerting social presence). In terms of the message's task engagement level, we find discussion forum contains a high proportion of learning-related messages (46.3% fully-engaged and 28.0% somewhat-engaged messages, vs. 25.7% disengaged). In comparison, in chat room, students posted more disengaged messages (60%) than fully-engaged messages (16.5%) and somewhat-engaged messages (23.5%).

4.2 Correlation Analysis

Because of the abnormal distribution of sense of community scores and activity features (see Tables 4 and 6), we use the Spearman correlation coefficient [36] to measure the relationship between students' communication behavior and their feeling of community. The results are given in Table 7.

4.2.1 Activity features and sense of community. It can be seen that there are significant correlations between students' activities in using communication tools and their sense of community. As for students' in-class activities, the number of messages students posted in chat room is significantly positively correlated with their sense of community w.r.t. both *Overall* scale and four sub-scales *Connectedness*, *Learning*, *Interaction with Instructor (InterInstructor)*, and *Interaction with Other Students (InterStudent)*. In addition, students' usage of hand-up facility has a significantly positive correlation with their sense of community in terms of *Learning* and *InterInstructor* sub-scales, indicating if students use hand-up facility more frequently, they are more likely to feel that the community can not only enhance their acquisition of knowledge and understanding (i.e., *Learning*), but also promote the interaction with their instructors (i.e., *InterInstructor*).

Regarding the after-class activities, it shows that when the number of learning materials students shared increases, their perception of *Learning* community also increases. In addition, students' assignment submission rate is significantly positively correlated with their sense of community w.r.t. *Learning* and *InterInstructor* sub-scales.

Thus, our results indicate that the usage of both synchronous online communication tools (like chat room and hand-up facility) and asynchronous communication tools (such as material sharing and assignment submission facilities) may enhance students' sense of community. In turn, strong sense of community may lead to students' increasing use of these communication tools. However, we fail to find significant correlations between students' behavior of taking notes, sharing notes, and posting messages in discussion forum, and their sense of community.

4.2.2 Content features and sense of community. In terms of messages' textual content, five features extracted from conversational texts of chat room show significant correlation with students' sense of community. More specifically, students who use larger number of social process words, affective process words, or cognitive process words in chat messages

Table 7: Correlations between implicit features and students' sense of community ($^{*}p < 0.05$ and $^{**}p < 0.01$)

			Sense of community				
			Overall	*Connectedness*	*Learning*	*InterInstructor*	*InterStudent*
Activity features							
In-class activities		Average # of using hand-up facility per lesson	0.084	0.057	**0.101***	**0.106***	0.070
		Total # of using hand-up facility	0.087	0.062	**0.102***	**0.108***	0.073
		Average # of chat messages per lesson	**0.169****	**0.133****	**0.171****	**0.176****	**0.138****
		Total # of chat messages	**0.168****	**0.134****	**0.168****	**0.175****	**0.136****
		Average # of taking notes per lesson	0.041	0.019	0.048	0.054	0.023
		Total # of taking notes	0.041	0.018	0.047	0.054	0.022
		Average # of sharing notes per lesson	0.011	0.021	0.005	0.010	0.014
		Total # of sharing notes	0.012	0.020	0.004	0.011	0.014
		Average course attendance rate	0.045	0.042	0.052	0.048	0.036
After-class activities		Total # of sharing materials	0.074	0.080	**0.098***	0.080	0.046
		Total # of forum messages	-0.017	-0.022	-0.008	-0.007	-0.033
		Average assignment submission rate	0.088	0.080	**0.097***	**0.124****	0.062
Content features							
Message length	Chat room	# of words per sentence	0.043	0.010	0.084	0.065	0.028
	Discussion forum	# of words per sentence	-0.018	-0.030	-0.012	-0.015	-0.032
Psycho-logical presence	Chat room	# of social process words	**0.177****	**0.141****	**0.182****	**0.188****	**0.149****
		# of affective process words	**0.175****	**0.140****	**0.187****	**0.189****	**0.143****
		# of cognitive process words	**0.174****	**0.123****	**0.195****	**0.187****	**0.144****
	Discussion forum	# of social process words	0.017	-0.02	0.042	0.034	-0.010
		# of affective process words	0.037	0.027	0.037	0.033	0.035
		# of cognitive process words	-0.012	-0.025	-0.003	-0.004	-0.028
Task engagement	Chat room	# of fully-engaged sentences	**0.136****	**0.092***	**0.154****	**0.153****	**0.099***
		# of somewhat-engaged sentences	**0.162****	**0.117****	**0.184****	**0.170****	**0.144****
		# of disengaged sentences	0.081	0.079	0.064	0.072	0.081
	Discussion forum	# of fully-engaged sentences	-0.045	-0.055	-0.011	-0.026	-0.054
		# of somewhat-engaged sentences	0.036	0.033	0.036	0.021	0.047
		# of disengaged sentences	-0.012	0.003	-0.031	0.002	-0.037
Miscellaneous features							
Personal properties		Gender	0.039	-0.006	0.072	0.055	0.022
		Age	-0.066	-0.088	-0.039	-0.038	-0.079
		Pre-course interest	**0.162****	**0.209****	**0.118****	**0.157****	**0.153****
		Pre-course knowledge	**0.102***	0.076	**0.131****	**0.106***	0.085
Course-related features		Course structure	-0.023	-0.024	-0.021	-0.016	-0.038
		Assessment structure	0.010	0.011	0.006	0.008	0.016
Environmental features^{+}		Instructors' aggregated activities	**0.101***	**0.089***	**0.108***	**0.118****	0.084
		classmates' aggregated activities	0.057	**0.090***	0.052	0.028	0.062

$^{+}$ Number of aggregated activities per lesson refers to the total number of activities carried out in each lesson. Specifically, for instructors, the activities include posting messages in chat room, taking notes, and sharing notes. While for students, the activities include using hand-up facility, posting messages in chat room, taking notes, and sharing notes.

are inclined to perceive stronger sense of community in terms of both *Overall* scale and four sub-scales, which implies the positive relationship between students' psychological presence and sense of community.

In addition, we observe that the numbers of chat messages which contain content fully or somewhat relevant to learning tasks are significantly correlated with students' sense of community (w.r.t. all of the scales) in a positive way. That is, when students post more fully-engaged or somewhat-engaged sentences in chat room, they are likely to perceive stronger sense of community.

However, there exists no significant correlation between the content features extracted from forum messages and students' sense of community in our study.

4.2.3 Miscellaneous features and sense of community. In addition to behavioral features, we find that some personal properties and environmental features are also associated with students' sense of community. Particularly, students who are more interested in the studied course (i.e., pre-course interest) or gain richer knowledge before class (i.e., pre-course knowledge) tend to show stronger sense of community, especially in terms of *Overall* scale and two sub-scales *Learning* and *InterInstructor*. The findings are basically consistent with Brown's observation [2] that students who are familiar with the online course tend to feel strong sense of community, because those who with less prior knowledge normally require more interaction with and support from online instructors.

Moreover, it shows that students' sense of community is related to their instructors' and classmates' behavior in synchronous online class. To be specific, the average number of instructors' aggregated in-class activities is significantly positively correlated with students' sense of community in terms of *Overall* scale and three sub-scales *Connectedness*, *Learning*, and *InterInstructor*, and the average number of classmates' aggregated in-class activities significantly correlates with students' feeling that they belong and connect to the online classroom community (i.e., *Connectedness* sub-scale).

However, our data reveal that students' age, gender, and course-related features (i.e., course structure and assessment structure) fail to show significant correlation with their sense of community ($p > 0.05$). Additionally, using analysis of variance (ANOVA), we notice that the mean differences of sense of community (w.r.t. both *Overall* scale and four sub-scales) across three subject types (i.e., liberal arts, science, and engineering) are not significant ($p > 0.05$).

In summary, 15 (out of totally 34) features are empirically proven to have significant correlation with students' sense of community values. Among them, there are more behavioral features (11 = 6 activity features + 5 content features) relative to miscellaneous features (4 = 2 personal properties + 2 environmental features). Particularly, students' activities carried out in class exhibit stronger correlation than their after-class activities. Regarding content features, the text

Table 8: RMSE results of testing regression models (Note: the models that significantly outperform the baseline are identified in bold, $p < 0.05$ via two-tailed paired t-test; and the value inside the parenthesis indicates the improvement percentage against the baseline approach)

	Baseline	LASSO	PR	Rule	GP	SVR	RBF
SC_Overall	0.1510	**0.1311* (13.2%)**	**0.1338* (11.4%)**	0.1479 (2.0%)	0.1456 (3.6%)	0.1471 (2.6%)	0.1428 (5.43%)
SC_Connectedness	0.1308	**0.1102* (15.8%)**	**0.1117* (14.6%)**	**0.1178* (9.9%)**	**0.1167* (10.8%)**	**0.1133* (13.4%)**	**0.1131* (13.5%)**
SC_Learning	0.1621	**0.1479* (8.8%)**	**0.1501* (7.4%)**	0.1593 (1.7%)	0.1557 (3.9%)	0.1560 (3.8%)	0.1567 (3.3%)
SC_InterInstructor	0.1570	**0.1403* (10.6%)**	**0.1416* (9.8%)**	**0.1421* (9.5%)**	**0.1441* (8.2%)**	0.1477 (5.9%)	0.1470 (6.4%)
SC_InterStudent	0.1412	**0.1219* (13.7%)**	**0.1228* (13.0%)**	**0.1258* (10.9%)**	**0.1262* (10.6%)**	**0.1286* (8.9%)**	**0.1237* (12.4%)**

contents of messages posted in chat room are more strongly correlated with sense of community, in comparison with contents extracted from forum messages.

4.3 Sense of Community Prediction

For the next step, we are interested in inferring the students' sense of community based on the 15 significant features identified in the previous section.

4.3.1 Inference model. Formally, a standard form of regression model can be represented as $y = f(x) + \epsilon$, where x denotes an input vector (in our case, it contains the identified features such as the number of posted chat messages, the number of social process words per message, etc.), y denotes a scalar output (in our case, it gives the predicted sense of community score), and ϵ is the additive noise. Our purpose is then to estimate the regression function $f(\cdot)$. In our experiment, we tested six popularly used regression models [7, 23, 33] for inferring students' sense of community: 3 linear methods including *Least Absolute Shrinkage and Selection Operator (LASSO)*, *Pace Regression (PR)*, and *M5 Rules (Rule)*, and 3 non-linear methods including *Gaussian Process (GP)*, *Support Vector Regression (SVR)*, and *Radial Basis Function Network (RBF)*.

Specifically, *LASSO* [26] is a shrinkage and selection method for linear regression to solve the following optimization puzzle: $\min_{\beta_0,\beta}(\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|)$, where N is the number of observations, y_i is the response at observation i, and x_i is a vector of p values at observation i. The parameters β_0 and β are the scalar and p-vector *LASSO* coefficients respectively. λ is the nonnegative weight given to the regularization term (the $L1$ norm). As for the regularization in *LASSO*, it is a powerful mathematical tool for reducing over-fitting, as it adds a penalty term to the objective function and controls the model complexity using that penalty term. *Pace Regression* [29] is a typical form of linear regression analysis, which improves on classical ordinary least squares regression by evaluating the effect of each variable. It is applicable when some of the input features are mutually dependent. *M5 Rules* [6] also assumes a linear distribution of the input features, but it is grounded on the separate-and-conquer strategy to build a decision tree. The advantage of *M5 Rules* is that it costs less calculation and can deal with the datasets with missing values.

As for non-linear models, *Gaussian Process* [16] defines a probabilistic regression based on Bayesian theory and statistical learning theory: $f(x) \sim gp(\mu(x), k(x, x'))$, where $\mu(x)$ stands for the mean function and $k(x, x')$ is the covariance function, which can handle datasets with small number of samples and/or many input features. As for the *Support Vector Regression* algorithm [24], the main idea is to minimize error and individualize the hyperplane which maximizes the margin. It maps the data into a high dimensional feature space via a nonlinear mapping and transforms the optimization problem into dual convex quadratic programs, which can get global optimum solution more efficiently. *Radial Basis Function Network* [3] is an artificial neural network having advantages of easy design, good generalization, and strong tolerance to input noise. It mainly uses radial basis functions as activation functions and the output of the network is a linear combination of radial basis functions of the inputs and neuron parameters.

4.3.2 Procedure. We randomly selected 80% of 489 students who participated in our user survey to train each model and tested it on the remaining 20% students. To avoid any biases, we performed 10-fold cross validation, and measured the accuracy via the metric *Root Mean Square Error (RMSE)*, which is a commonly used measure of the difference between predicted value and ground truth (the lower, the better) [31]. All significance tests were done using two-tailed paired t-test at the $p < 0.05$ level. Formally, we define a student's sense of community as a 5-dimension vector $sc_u = (sc_u^1, sc_u^2, ..., sc_u^5)^T$, where sc_u^1 represents the *Overall* scale and sc_u^2 to sc_u^5 respectively represent the four sub-scales *Connectedness*, *Learning*, *InterInstructor*, and *InterStudent*. The means and standard deviations of the five dimensions at the normalized 0-1 scale are: *Overall* (mean=0.53, st.d.=0.15), *Connectedness* (mean=0.57, st.d.=0.13), *Learning* (mean=0.49, st.d.=0.16), *InterInstructor* (mean=0.56, st.d.=0.16), and *InterStudent* (mean=0.53, st.d.=0.14).

4.3.3 Prediction results. The results are shown in Table 8. We observe that the six regression models all achieve significant improvements against the baseline that simply uses the average value of training data as the predicted score for all of the testing samples ($p < 0.05$), among which the linear methods *LASSO* and *Pace Regression* perform better than non-linear methods. *LASSO* is further better than *Pace Regression* in terms of all the five dimensions (average RMSE value: 0.1303 vs. 0.1320). It is probably because LASSO has the ability to deal with the over-fitting problem that may occur in our dataset (more features and fewer samples). In contrast, some methods like *SVR* and *RBF Network* fail to reach satisfying prediction accuracy, which may be due to the linear characteristics of our data that do not fit their non-linear assumption.

In addition to the RMSE values, we also report the improvement percentage ($= \frac{|RMSE_{testmodel} - RMSE_{Baseline}|}{RMSE_{Baseline}}$) that each model achieves against the baseline. It shows that the improvement percentage returned by LASSO w.r.t. the *Overall* scale is 13.2%. As for the four sub-scales of sense of community, *Connectedness* is the easiest one inferred by LASSO (15.8% accuracy increase relative to the baseline), followed by *InterStudent* (13.7%), *InterInstructor* (10.6%), and *Learning* (8.8%), implying that our identified features are more effective at reflecting students' feeling of connectedness with community members.

5 DISCUSSION

5.1 Major Findings

In our work, we not only reveal the correlation between students' communication behavior in online courses and their sense of community, but also build regression model for inferring students' sense of community based on their behavior.

First of all, we find both students' activities (6 features) and message contents (5 features) are significantly correlated with their sense of community. In synchronous class, we observe that students' sense of community increases as their usage of chat room or hand-up facility rises. This may be because the immediate and direct interaction can make students feel more connected to the community and achieve their learning goals. In turn, feeling stronger sense of community may encourage students to behave more actively in using these communication tools. After class, we find that students' usage of material-sharing and assignment-submission facilities are significantly correlated with their sense of community in terms of the *Learning* sub-scale. As for content features, the psychological presence and task engagement level of the messages posted in chat room show significant correlation with students' sense of community. In addition, students' pre-course interest, pre-course knowledge, and the average numbers of instructors' and classmates' aggregated in-class activities are also significantly postively correlated with students' sense of community. However, we fail to observe any significant findings regarding students' behavior in discussion forum, which is probably because the delayed communication may make students feel less motivated to interact.

Motivated by the correlation results, we further developed inference model to identify students' sense of community based on these significant features. Concretely, we tested six machine learning algorithms including *LASSO*, *Pace Regression*, *M5 Rules*, *Gaussian Process*, *Supported Vector Regression*, and *Radial Basis Function Network*. Our results demonstrate that all of the models significantly outperform the baseline, and *LASSO* performs the best. The possible reason is that *LASSO* is capable of alleviating the over-fitting issue, and the linear relationship between input and output as defined in this model may better fit the characteristic of our data. Another observation is that *Connectedness* and *InterStudent* sub-scales are easier to be predicted relative to *InterInstructor* and *Learning* sub-scales, probably because students' feeling of connectedness with community members and their perception of interaction with learning peers can be better reflected by the significant features.

5.2 Implications

Therefore, we believe that our results are suggestive for researchers to better understand the relationship between students' usage of different communication tools and their sense of community in online courses, as well as for practitioners to improve existing online learning systems. For instance, more synchronous tools such as chat room and hand-up facility may be incorporated into current products, so as to enhance students' connectedness and interaction with community members. In addition, instructors could increase their initiatives in using these communication tools in synchronous class, as their in-class behavior is positively correlated with students' sense of community. Instructors could also assign more homework due to the positive relationship between sense of community and assignment submission.

Furthermore, the ability to infer a student's sense of community from her/his online communication behavior could be potentially helpful to address the dropout issues in current online courses. As instructors would be able to know their students' sense of community in real time, once the value degrades below an acceptable threshold, they may offer pertinent and timely supports to their students. For instance, when it shows students perceive lower sense of community in terms of *Learning* and *InterInstructor* sub-scales, instructors could ask more questions in class so as to encourage students to answer by using hand-up facility, while for those who feel lower sense of community regarding *Connectedness* and *InterStudent* sub-scales, more chat sessions could be organized for narrowing the distance between learning peers.

6 CONCLUSION AND FUTURE WORK

Although students' sense of community plays an essential role in online learning, how to acquire it in real time remains a big concern. Our study suggests that it is feasible to infer students' sense of community from their communication behavior in online courses. To be specific, we first identified a set of features which are significantly correlated to sense of community, not only including students' activities carried out in both synchronous and asynchronous environments, but also their linguistic content in conversational texts. We then compared six regression models in terms of their ability of unifying these features into automatically predicting students' sense of community, among which *LASSO* shows the best performance.

Our work has several future directions. Firstly, we plan to validate our findings on more students with diverse backgrounds (e.g., age, nationality, ethnic background). Secondly, we will try to further improve our prediction model by considering more features, such as the semantic content of message texts. We will also perform qualitative interviews to in depth understand students' thoughts.

ACKNOWLEDGMENTS

We thank all participants who took part in our user survey. We also thank Hong Kong Research Grants Council (RGC) for sponsoring the research work (under project RGC/HKBU12200415).

REFERENCES

[1] Mohamed Ally. 2004. Foundations of educational theory for online learning. *Theory and Practice of Online Learning* 2 (2004), 15–44.

[2] Ruth E Brown. 2001. The process of community-building in distance learning classes. *Journal of Asynchronous Learning Networks* 5, 2 (2001), 18–35.

[3] Sheng Chen, Colin FN Cowan, and Peter M Grant. 1991. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks* 2, 2 (1991), 302–309.

[4] Sue-Jen Chen and Edward J Caropreso. 2004. Influence of personality on online discussion. *Journal of Interactive Online Learning* 3, 2 (2004), 1–17.

[5] Shane Dawson. 2006. Online forum discussion interactions as an indicator of student community. *Australasian Journal of Educational Technology* 22, 4 (2006), 495–510.

[6] Geoffrey Holmes, Mark Hall, and Eibe Prank. 1999. Generating rule sets from model trees. In *Australasian Joint Conference on Artificial Intelligence (AI 1999)*. Springer, 1–12.

[7] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* (2007), 457–500.

[8] Joanne M McInnerney and Tim S Roberts. 2004. Online learning: Social interaction and the creation of a sense of community. *Educational Technology & Society* 7, 3 (2004), 73–81.

[9] John Paul McKinney, Kathleen G McKinney, Renae Franiuk, and John Schweitzer. 2006. The college classroom as a community: Impact on student attitudes and learning. *College Teaching* 54, 3 (2006), 281–284.

[10] David W McMillan and David M Chavis. 1986. Sense of community: A definition and theory. *Journal of Community Psychology* 14, 1 (1986), 6–23.

[11] Robert L Moore. 2014. Importance of developing community in distance education courses. *TechTrends* 58, 2 (2014), 20–24.

[12] Jum C Nunnally, Ira H Bernstein, and Jos MF ten Berge. 1967. *Psychometric Theory*. Vol. 226. JSTOR.

[13] Murat Oztok, Daniel Zingaro, Clare Brett, and Jim Hewitt. 2013. Exploring asynchronous and synchronous tool use in online courses. *Computers & Education* 60, 1 (2013), 87–94.

[14] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works* (2015).

[15] Krystal Phirangee, Carrie Demmans Epp, and Jim Hewitt. 2016. Exploring the relationships between facilitation methods, students' sense of community, and their online behaviors. *Online Learning Consortium* (2016).

[16] Carl Edward Rasmussen. 2006. Gaussian processes for machine learning. *Citeseer* (2006).

[17] Ido Roll, Leah P Macfadyen, and Debra Sandilands. 2015. Evaluating the Relationship Between Course Structure, Learner Activity, and Perceived Value of Online Courses. In *Proceedings of the 2nd ACM Conference on Learning@ Scale (L@S 2015)*. ACM, 385–388.

[18] Alfred P Rovai. 2002. Development of an instrument to measure classroom community. *The Internet and Higher Education* 5, 3 (2002), 197–211.

[19] Alfred P Rovai. 2002. Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. *The Internet and Higher Education* 5, 4 (2002), 319–332.

[20] Alfred P Rovai and Hope Jordan. 2004. Blended learning and sense of community: A comparative analysis with traditional and fully online graduate courses. *The International Review of Research in Open and Distributed Learning* 5, 2 (2004).

[21] Seymour B Sarason. 1974. *The psychological sense of community: Prospects for a community psychology.* Jossey-Bass.

[22] Sarah Schrire. 2006. Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis. *Computers & Education* 46, 1 (2006), 49–70.

[23] Peter Shea. 2006. A study of students' sense of learning community in online environments. *Journal of Asynchronous Learning Networks* 10, 1 (2006), 35–44.

[24] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 3 (2004), 199–222.

[25] Malin Sveningsson. 2001. *Creating a sense of community: Experiences from a Swedish web chat.* Ph.D. Dissertation. Linköpings universitet.

[26] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[27] Jesus Trespalacios and Ross Perkins. 2016. Sense of Community, Perceived Learning, and Achievement Relationships in an Online Graduate Course. *Turkish Online Journal of Distance Education* (2016).

[28] Donald G Unger and Abraham Wandersman. 1985. The importance of neighbors: The social, cognitive, and affective components of neighboring. *American Journal of Community Psychology* 13, 2 (1985), 139–169.

[29] Yong Wang. 2000. *A new approach to fitting linear models in high dimensional spaces.* Ph.D. Dissertation. The University of Waikato.

[30] Rupert Wegerif. 1998. The social dimension of asynchronous learning networks. *Journal of Asynchronous Learning Networks* 2, 1 (1998), 34–49.

[31] Cort J Willmott, Steven G Ackleson, Robert E Davis, Johannes J Feddema, Katherine M Klink, David R Legates, James O'donnell, and Clinton M Rowe. 1985. Statistics for the evaluation and comparison of models. *American Geophysical Union* (1985).

[32] Robert H Woods Jr. 2002. How much communication is enough in online courses?–exploring the relationship between frequency of instructor-initiated personal email and learners' perceptions of and participation in online learning. *International Journal of Instructional Media* 29, 4 (2002), 377.

[33] Wen Wu and Li Chen. 2015. Implicit acquisition of user personality for augmenting movie recommendations. In *International Conference on User Modeling, Adaptation, and Personalization (UMAP 2015)*. Springer, 302–314.

[34] Wen Wu, Li Chen, and Qingchang Yang. 2016. Students' Personality and Chat Room Behavior in Synchronous Online Learning. In *Proceedings of ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2016), Late-Breaking Results*. ACM.

[35] J Yuan and C Kim. 2014. Guidelines for facilitating the development of learning communities in online courses. *Journal of Computer Assisted Learning* 30, 3 (2014), 220–232.

[36] Jerrold H Zar. 1972. Significance testing of the Spearman rank correlation coefficient. *J. Amer. Statist. Assoc.* 67, 339 (1972), 578–580.

[37] Saijing Zheng, Mary Beth Rosson, Patrick C Shih, and John M Carroll. 2015. Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW 2015)*. ACM, 1882–1895.