ABSTRACT

XU, YIQIAO. Analyzing and Scaffolding Semi-Structured MOOC Forums to Support Social Learning and Help-Seeking. (Under the direction of Dr. Collin F. Lynch.)

Massive Open Online Courses (MOOCs) are purely online learning environments that are typically available for free for anyone with internet access. These courses have attracted hundreds of thousands, if not millions of learners around the world and the number of users has increased rapidly along with the growth of new platforms and a wide diversity of courses. MOOC classrooms are typically structured around a common set of educational tools including a basic learning management system, video lectures, quizzes, and a discussion forum. In MOOC, discussion forum is the only medium for student communication, which records a rich amount of massive data on students' activities. Via students forum interaction, we can understand what their learning behaviours are and how these learning behaviours impact their course performance. Understanding students' behaviours can also help instructors provide useful interventions and improve students' learning gain.

In this work, we analyze students' discussion forum activities in several offerings of one MOOC which was delivered on two different platforms. We collected students interactions from the course platforms and the discussion forum to explore their social communication and learning behaviours. To organize the semi-structured forum, We first examine how MOOC students build their social community. Then we model students' learning behaviours based on their discussion forum post content. Eventually, we automatic generate interventions for instructor and suggestions for students based on analyzing students post content.

Organizing semi-structured MOOC discussions can help us to improve students' educational experiences in many ways. First, by analyzing student weekly online activities, we track evolution of study over time by topic modeling algorithm. Second, via language analysis on posts content, we provide instructors with a dashboard that shows the evolution of topic over time. Last, we respond to posts (particularly questions) with pointers to course materials and current course posts that are relevant. The results of our work can provide instructors with insights into students online behaviours from the semi-structured discussion forum and also provide students supports to improve their learning gain.

Analyzing and Scaffolding Semi-Structured MOOC Forums to Support Social Learning and Help-Seeking

by
Yiqiao Xu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Physics

Raleigh, North Carolina

2022

APPROVED BY:

Tiffany Barnes	Sarah Heckman
Cesar Delgado	Dr. Collin F. Lynch Chair of Advisory Committee

DEDICATION

This dissertation is dedicated to all the people who have impacted my Ph.D. journey.

To my Ph.D. advisor Dr. Collin Lynch who brought me to the exciting fields, who has not only supported my research work but also guided my life.

To Dr. Min Chi, who gave my family a lot of guidance when my son born.

To my beloved wife Yao Luo who supports me through the entire Ph.D, career.

To my little cute son Anze who brought me the incredible love.

To my parents Yunhai and Li, who bring life to me, encourage and help me through my entire life.

BIOGRAPHY

Yiqiao Xu was born in Jinan, China. He was introduced to computer science programming during middle school and achieved a gold model in the Chinese national Lego robotics competition. He joined Xi'an Jiaotong University in 2011. And he graduated from the undergraduate program of Automation at Xi'an Jiaotong University and completed all the courses with qualified academic standing in 2015. He is conferred upon Bachelor degree of Engineering.

In 2015, he moved to Raleigh, NC, and joined the Department of Computer Science at North Carolina State University as a Ph.D. student. In the fall semester of 2016, he dived into educational data mining research. He has worked as a research assistant in ArgLab under the supervision of Dr. Collin Lynch. He was involved in a research project analyzing students' discussion forum help-seeking behaviours. For the first part of his research career, he has donated his time to MOOC students' help-seeking behaviour analysis. Later, he investigated his research into the instructor-based study to apply machine learning algorithms to support instructor teaching and improve students learning experience. He also participated in designing a platform, 'CONCERT', which collects students' data from different sources, prepares for analysis and visualizes for instructors. During his Ph.D. years, he worked with Meta doing internship to get more familiar with the industry work. He was also involved in the Educational Data Mining and Innovative Use of NLP for Building Educational Applications communities by reviewing conference papers, volunteering at the conferences, and helping with conference organizations. In June 2022, he defended his dissertation with the title 'Analyzing and scaffolding semi-structured MOOC forums to support social learning and help-seeking and joined Metlife as a data scientist afterward.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Collin Lynch for all his support during my Ph.Dd years. He helped my get involved in the research domain and the community.

I would also like to thank my committee members, Dr. Tiffany Barnes, Dr. Sarah Heckman, and Dr. Cesar Delgado. They helped my polished my research works and always made out times to review my papers.

Many thanks to my friends at ARG Lab: Linting Xue, Niki Gitinabard, Adam Gaweda, Jennifer Tsan, Zhikai Gao, and Brad Erickson for all discussions and collaborations. I appreciate you all.

Finally, I would like to thank my wife, Yao Luo, for being my partner, my family, and mother of my son. It would be impossible to get through the challenges of graduate school and live away from home without her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Massive Open Online Courses (MOOCs) hold the potential to open up educational opportunities to a global audience with a low cost. The broad goal of MOOCs is to scale instruction by allowing expert instructors to provide guidance to hundreds or even thousands of students at a time. Such high volume education has the potential to be revolutionary both for individual students and educational systems. The current generation of MOOCs are designed to achieve this scaling by outsourcing much of the individual support tasks to students. That is, rather than capping enrollment to ensure that the instructor and TAs can support every students' needs, MOOCs provide online forums that encourage students to share common questions and to provide collaborative guidance or to benefit from each others' interactions with the limited support staff. Thus it is tacitly assumed that students will have common issues and that good students will help poor students with course content, assignments, logistics, and other issues.

Social network analysis has been used for a long time to study online communities such as health communities [Ste11], popular question answer forums [Ada08], and internet relay chat (IRC) [Pak11; Ros09; Ros11; Mac]. A number of prior studies have investigated the benefits of collaborate learning. MOOC discussion forums offer a unique medium for collaborative learning activities where students and their peers share their opinion and feedback. Thus, there is a theoretical emphasis in the literature on the role of discussion forums for collaborative learning activities [Cal10]. From the teachers' perspective, these type of learning activities provide insights into students' learning processes [Rab14]. Teachers are able to know what has been learned and what challenges student faced during the course by drawing insights from the forum. By working on the course online independently, MOOC students engage in a self/peer social process that help them to complete the course. However, based upon the characteristics of MOOCs, a large number of posts are generated within in a few weeks in the forum, which also contain lengthy discussions bearing many interactions between students. The amount of forum data could be overwhelming to MOOC instructors to manually read through. Thus, many prior studies analyze MOOC students' behaviours by social network analysis based upon the reply relation in the discussion forum with data mining technologies. Sinha et al. [Sin14a] analyzed an ongoing online course offered in Coursera using a social network analysis to identify students who are actively engaging in the course and those who are potentially at risk of dropping out. They found participation patterns of distinct groups of students in the networked learning community and effectively discovered important discussion threads. Brown et al. investigated the how MOOC students' social communities impacted their course performance measured by their final grade [Bro15a]. They found that students will often communicate with similarly-performing students instead of helping different students in different groups. However, most of these prior studies were limited by the fact that they only used the final social network from courses. Thus, in order to provide the guidance and support students learning, we extended the work of Brown et al. [Bro15a] by investigating the process of how students form their social community during the course. Prior research has shown that students form stable social structures in MOOCs, however the impact of those connections on students' performance has not always been consistent. Jiang et al. analyzed an algebra MOOC, and found that attributes of the students' social network were correlated with the students' final grades, but they found no relationship between the same variables in a different MOOC on finance [Jia14]. Houston et al. likewise examined the forum activities that were most strongly associated with final grades in three different MOOCs and found that the addition of social centrality and similar features had no impact on their predictive models [HI17]. This inconsistency may be explained by the fact that the types of discussions students have and their relevance can change from class to class and that prior analyses have focused primarily on the overall social structure and not the content of the discussions.

While peer communication features in MOOCs are intended to foster content engagement, many of the most active discussion-topics are often social conversations, critiques of the class videos, or exchanges of career advice [Sea14]. Prior researchers have developed automated detectors to classify these posts into on- and off-topic comments and to evaluate the relative proportion of relevant discussions to learning outcomes [Lin09; Wis16]. However, while they have devoted a great deal of effort to improve these models relative to expert labels, they have not connected these discussions to students' social networks in the course to understand the distribution of the discussions, evaluate the relative impact of on- and off-topic posts on students' engagement, or identify a *pragmatic baseline* where the models, however imperfect, are sufficient for class use. We explored how the topical content of forum discussions affects students' social relationships, as well as how they connect to their learning outcomes.

On the other hand, MOOCs are purely online learning environments that are typically available for free to anyone with internet access. These courses have attracted hundreds of thousands of people to register and the number of users have increased rapidly due to the growth of platforms and the wide range of offered courses [Kay13; McA10].

MOOC classrooms are typically composed of a basic learning management system, video lectures, quizzes, and a discussion forum [Shi15]. In this work, Shi et al. developed VisMOOC, a visual analytic system to help analyze user learning behaviors by using clickstream data from MOOC platforms. It provides three main views: the List View to show an overview of the clickstream differences among course videos, the Content Base View to show temporal variations in the total number of each type of click action along the video timeline, and the Dashboard View to show various statistics such as demographic and temporal information. However, in light of the fact that MOOCs attract large numbers of people to register and engage in the discussion forum, students generate a large volume of clickstream data by using these common educational tools. These tools provide insights into their learning behaviours, engagement activities, and their learning habits.

Further, learning in MOOCs requires students to apply self-regulation. Prior studies have analyzed clickstream data to investigate MOOC student learning status [He15; Bal13; Boy15; Hal14a]. For example, Brinton et al. [Bri15] studied students performance prediction in MOOCs where the objective was to predict whether a user will be Correct on First Attempt(CFA) in answering a question. They define Correct on First Attempt as the correctness when MOOC students first attempt their quizzes. From the clickstream data, they extracted information such as fraction of lectures played, number of pauses for each user-video pair, and show how CFA reflects students learning performance. Sinha et al. [Sin14b] explored students' learning experience in MOOC through video lecture interaction. Their results illustrate how such features help understanding students' course engagement. Based upon their results, they were able to use students' clickstream interactions to predict student dropout. Thus, video lectures form a primary part of MOOC instructor design. Concept discussion and tutorials that are included in the video lectures not only guide the students to finish their assignments in order to complete the course but also encourage students to participate in the discussion forums.

Thus, organizing semi-structured MOOC discussion forums can help us to understand students' online learning behaviours. We can use our findings of how students established their social community to guide them seeking help from their peers. We can also use the content analysis of forum posts to automatically monitor student discussion topics over time and generate material recommendations to students' questions. In this work, my goal is provide methods to organize MOOC discussion forums and provide support to instructors and students based upon my findings. I will answering the following research questions:

- RQ1: How can we model the communicative activities of MOOC students based on their forum posts?
  - RQ1.1: How do students build their communication over time based on their forum reply relationship?
  - RQ1.2: How do we understand students forum post by categorizing the post content?
  - RQ1.3: How do we analyze student online activity transactions combined with their clickstream and forum activities?
- RQ2: How can instructors monitor students' learning status based on their use of the forum? And can we use our knowledge of structured forums to help instructors make initial interventions and improve students' learning experience and get answers more efficiently?

  - By efficiently partitioning requests for help from other contexts.
  - By clustering similar questions to recognize duplicate and non-answered topics.
  - By aligning questions to related course materials.

- RQ3: What are instructors' goals in monitoring student interactions? How do they evaluate adaptive scaffolds? What features do they want in an adaptive scaffold for content monitoring?

Answering RQ1 will help researchers to better understand MOOC students, their online learning behaviours, their social connections, and their help seeking. This analysis can make the online learning behaviour of better performing students more clear, which makes it easier for instructors to learn how MOOC students build their communication and seek help from each other. Exploring the second research question will help us to understand students' learning processes from massive discussion forums. With this finding, instructors are able to learn what students discuss and how they engaged in reflective thought from their forum post content. Then we develop a dashboard intervention to help instructors monitor students' learning process. Finally, investigating the third research question investigates what are instructors' goals in monitoring students and what feature they want to support their teaching by monitoring students from the discussion forum.

In Chapter 2, I will discuss prior studies in relevant areas. I will describe the dataset I used for this work in Chapter 3. The detailed background, methodology, findings, and conclusion of Research Question 1 will appear in Chapters 4 - 6. Research Question 2 will be discussed in Chapter 7. Research Question 3 will be described in Chapter 8. Finally, I will discuss my conclusions, future work, and limitations in Chapter 9.

Chapter 2

BACKGROUND

2.1 Social Learning Theory

Learning is often a social process and the social aspects of learning have been analyzed by researchers in the past decades. Reed et al. [Ree10] provided a definition of social learning. They argued that to be considered social learning, a process must (1) demonstrate that a change in understanding has taken place in the individuals involved; (2) demonstrate that this change goes beyond the individual and becomes situated within wider social units or communities of practice; and (3) occur through social interactions and processes between actors within a social network. Laland et al. [Lal04] explored the strategies of social learning. In this article, they identified a number of possible strategies that are predicted by theoretical analyses, including 'copy when uncertain', 'copy the majority', and 'copy if better', and consider the empirical evidence in support for each, drawing from both the animal and human social learning literature. Shum et al. [Shu12] propose that the design and implementation of effective Social Learning Analytics (SLA) present significant challenges and opportunities for both research and enterprise, in three important respects. They posit Social Learning Analytics as a response to some of the forces reshaping the educational landscape, and our growing understanding that many forms of learning are socially grounded and evidenced phenomena. SLAs may be deployed as institutional tools in conventional courses to yield insights for educators and administrators. Equally, they should be seen as tools for the very subjects being analysed—the learners—and for the many informal learning contexts that we now see outside of conventional institutions. It would indeed be ironic if the ways in which Social Learning Analytics tools were deployed did not honour and promote the open, democratising, critical dynamics that underpin much of the participatory, social web philosophy—dynamics which SLA tools make visible in new ways.

On the other hand, Merriam and Caffarella [Caf99] define constructivism as the assimilation of both behaviourist and cognitive ideals. In their words: "constructivist stance maintains that learning is a process of constructing meaning; it is how people make sense of their experience". Mvududu et al. [Mvu12] claimed that constructivism is widely known as an approach to probe for children's level of understanding knowledge and to show how that kind of understanding changes their high level of thinking. Thus, with constructivism as an educational theory, teachers should consider their students' knowledge and encourage them to put it into practices.

Piaget [Pia77] developed the framework for cognitive constructivism. He asserted that learning does not occur passively, instead it occurs by active construction of meaning. He argues that when learners encounter an experience that challenges the way they think, a state of imbalance is created. Learners have to change their thinking to restore balance. However, Lev Vygotsky's work [Vyg78] is critical upon Piaget's contribution to constructivism. While Piaget believes that development precedes learning, Vygotsky believes the opposite. On the topic of the development of speech, Piaget said that the children's egocentric speech goes away with maturity and is the transformed in to social speech. On the contrary, Vygotsky stated that childrens' minds are inherently social in nature and so speech moves from communicative social to inner egocentric. Therefore, since the development of thought follows the development of speech, Vygotsky claims that thought develops from society to the individual and not the other way.

Thus, social learning plays an important role in student learning processes. In most of the prior work, the researchers believed that in order to help the students learn from their environment, it is better to sit them into a psychical learning environment such as classroom. However, MOOC is an unique classroom environment rapidly developed in last decade. For MOOCs the only environment for students to have social interaction is the discussion forum. Students post questions and answers in the forum to discuss the course content with their peers and instructional staff. In addition, students' activities in the discussion forum represent their learning process and social behaviours. Thus, we believed that these features could be used to understand students' learning behaviours.

2.2 Student Help Seeking

The concept of self-regulated learning has gained increased recognition in the last decade as being important for student learning and achievement. Self-regulated learners are students who are metacognitively, motivationally, and behaviorally involved in their learning [Zim94]. One specific characteristic of a self-regulated learner in the classroom is the ability to use others as a resource to copy with ambiguity and difficulty in the learning process. It is inevitable that students will encounter a situation in which they need aid or advice to continue an academic task. In such a situation, a student must become aware of needing help (metacognition), decide to seek help, and implement strategies for engaging another person's help.

Newman [New90; New94] proposed that when students decide to seek help their decision is filtered by a motivational affective system that includes students' perception of competence, achievement goals, and attitudes. Based on that theory, Ryan et. al. [Rya97] investigated how these differences influence their help seeking. Specifically, this study focused on how students' perceptions of competence and achievement goals were related to their attitudes toward help seeking and their self-reported help-seeking behaviors in a math class. They observed two help-seeking behaviors: avoidance of help seeking and adaptive help seeking. Avoidance of help seeking refers to instances when a student needs help but does not seek it. For example, a student might skip a problem altogether or put down any answer rather than ask for help. When students don't garner help when it is needed they put themselves at a disadvantage for learning and performance. In contrast, adaptive help seeking involves a student asking for hints about the solution to a problem, examples of similar problems, or clarifications of the problem. Such help-seeking strategies are adaptive in that the help requested is limited to the help needed by the individual to solve the problem independently.

As these studies show, help-seeking is an important behaviour for students to improve their learning. In most of the previous work, the researchers concluded that for students to seek help to solve problems as a community, they should have in-person interactions. However, there is some debate about how well the theory fits into the online learning environment, i.e., discussion forums for MOOCs. The online discussion forum is the only venue for participants to communicate in a MOOC environment. Thus, we analyzed students' online forum activity patterns to understand how different performing students seek help from their peers.

2.3 Students' online learning behaviours

Social connection is a very important aspect for students learning from each other. Over one-third of teaching faculty in US use social media in their courses and adoption rates of social media are as high as 80% in university classrooms [Mor11]. Another study investigated the technologies that matter most to undergraduate students by exploring students' technology experiences and expectations [Dah14]. They sent out a survey to 1.5 million students at 213 institutions across 15 countries, yielding 75,306 responses. Their findings indicated that social media was being formally integrated into institutional academic learning experience and used by students to support their learning experience. Ito et al. [Ito13] investigated how instructors can use social media to support connected learning in a broader way. Social media addressed the gap between in-school and out-of-school learning by enabling the discovery of connections between their traditional curricula, their personal interests, and online communities that can support and further their engagement and learning.

2.3.1 Social Network Analysis

Studies in educational social network analysis are supported by rich information as social media are used to form and maintain social connections and share information from peers. These connections describe the social networks of who is communicating with whom, what they are discussing, and how they exchange their knowledge information. There has been prior research analyzing students' social network and understanding the information to support students' learning. Gruzd et al. [Gru16] discussed methods that can help educators to evaluate and understand the observed and potential use of social media for teaching and learning through social network analysis. Anaya et al. [Ana15] investigated how MOOC students collaborated online and how that collaboration influenced their learning. They proposed an approach to automatically inform teachers on students' problematic collaboration so that teachers can provide suggestions to students when necessary.

On the other hand, Boroujeni et al. [Bor17] investigate the dynamics in MOOC discussion forums by analyzing the interplay of temporal patterns, discussion content, and the social structure emerging from communication. They revealed the dependencies between the course structure, video openings, assignment deadlines, and the overall activity from the online forums. They applied modeling techniques from social network analysis to analyze the social dimension. While the types of user roles based on connection patterns are relatively stable over time, the high fluctuation of active contributors lead to frequent changes from active to passive roles during the course. However, while most users do not create many social connections they can play an important role in the content dimension triggering discussions on the course subject. Finally, they found that forum activity level can be predicted one week in advance based on the course structure, students' forum activity history, and attributes of the communication network which enables identification of periods when increased tutor support in the forum is necessary. These categories are similar to the ones suggested by Kizilcec et al [Kiz13]. While clustering seems to offer much insight on similar sequences and differences between different groups of students, it is often challenging to interpret these clusters and apply this knowledge to real world groups of students.

In addition, Gitinabard et al. [Git18] conducted a survival analysis to identify the likelihood of MOOC student dropouts with behavioral and social features. They built a social network from the discussion forum and then they applied feature selection methods to find out which features can provide instructors with more learning gain on students dropout. Finally, they used a prediction model trained on each week of an earlier offering of the course to predict dropout and certificate earning on another offering of the course. They found out that behavioral features such as assignments submissions and video watching are better predictors of student dropout and certification than social features.

2.3.2 Content Analysis in MOOCs

Social media produces a large amount of textual data that records the history of group interaction as network and communities establish. Content analysis is a technique to analyze the pattern of text and language. Text based sentiment analysis relies on natural language processing to compress large amounts of text data into patterns and categories that enable researchers to investigate [Kri18]. When it comes to educational contexts, content analysis has been used to identify exploratory dialogue [EC13; Fer13], and apply topic modeling to predict student dropout [Ram14a]. In addition to making use of the social network matrix (centrality, in-/out-degree) and social network communities, some studies, e.g. [Wan15a; Wan16a; Wis18], have also incorporated forum content analysis into their models of students' contributions. Wang et al. applied a content analysis approach to analyze students' behaviours on MOOC forums and explored how participation influenced their learning gains [Wan15a]. They modeled the participation using the Interactive, Constructive, Active, & Passive (ICAP) framework, as well as classified each post into the four categories of the ICAP framework.

Wang et al. observed that students' *active* and *constructive* discussion activities were highly relevant to predicting their learning performance. Constructive behaviours are defined as when the learner produces output, which could be examples or explanations that go beyond course materials. According to the ICAP framework, constructive activities should provide better learning outcome than active activities. However, they didn't observe this pattern from their study. Their explanation is that their post-test may not have targeted the skills and concepts students learned from these constructive activities. A follow up study further explored what kinds of discussion forum behaviours contributed to higher learning gains [Wan16a]. They found that students who exhibited more high-order thinking behaviours received better grades. They also distinguished socially-oriented topics as 'social chat' and biopsychology-oriented topics as 'biopsychology related topics' in a psychology course. They also applied a topic modelling method on the course materials to examine the relationship between topics and student activities. They concluded that socially-oriented topics are positively associated with richer discussions, which means more discussions on the forum, while biopsychology-oriented topics are negatively associated with richer discussions. Similarly, Wise et al. analyzed how social matrices based upon the content of the discussion posts effected student learning [Wis18]. In their study, they distinguished discussion between course related and non-course related posts by classifying the start of thread posts (accuracy ¿ 0.81). They found that students with both course related and non-related posts were more likely to pass the course. However, of those who did pass the course, there was no significant difference between the scores of the related and non-related groups, though students that posted course-related topics earned slightly higher final scores than their peers.

2.3.3 Forum Topic Detection

Detecting topics from online forums is an efficient method to monitor the state of the forum to identify emerging trends. Topic modeling has been wildly used to identify issues and people's opinions. Researchers conducted studies to identify patterns of different professional domains such as health, education, finance, etc, from different kinds of data sources, communities, and locations. On the other hand, LDA (Latent Dirichlet allocation) [Ble03], a probabilistic topic model used to uncover latent topics in texts, has been widely used to extract topics from different data sources.

In addition, in educational context, topic modeling has been used to summarize MOOC posts based on the inferred topics. Thushari et al. [Ata16a] proposed a framework to automatically generate and label MOOC discussion posts based on topic modeling. The authors extended the work to propose a topic visualization dashboard that can aid in understanding emerging discussion themes and identifying popular topics of discussion in MOOC forums [Ata16b]. They claimed that their topic modeling dashboard allowed the instructors to locate the most influential discussion topics. This allowed the instructors to respond to those topics based on the corresponding course material. Automatically classifying the topic of discussion can aid in providing adaptive support to individual students and to collaborative groups. Song et al. [Son17] proposed a data-driven approach called TOLA (Topic-oriented learning assistance) to study students' learning patterns using LDA to extract topic features which integrated hybrid features such as the number of posts and the number of participants.

Predicting student survival and social network analysis of MOOC participants is another important subject matter where LDA has been applied. Ramesh et al. [Ram14b] proposed a model to predict student survival in online courses using linguistic features and Seeded LDA to categorize discussion posts and behavioral features such as lecture views and posting/voting/viewing discussion forum content. Yang et al. [Yan14] used LDA based analysis of discussion forum text and social network features to study the emergence of student sub-communities and predict student drop-outs in MOOCs. Wang et al. [Wan16b] studied the relationship between student performance and higher-order thinking behavior in MOOCs using LDA to identify topics associated with the occurrence of higher-order thinking behaviors. Brinton et al. [Bri16] applied LDA to extract topics from online course forum data to study topic-wise collaboration and dissemination tendencies among students and investigated the efficiency of Social Learning Networks (SLN). Robinson et al. [Rob15] used LDA on a Cartography course offered on Coursera to investigate the global reach of MOOCs and the geography associated with place names mentioned in the posts by identifying key themes and geographic references in the discussion posts.

Prior researchers analyzed how students' online forum activities impact their learning. However, most of them didn't develop a general consistent structure to analysis how can instructors learn from the student forums and what can instructors do to improve students' learning experiences based on their forum features. Different from previous studies, we examine what students do in the forum and structure their activities to investigate how we can assist instructors to help students improve their learning performance.

2.4 Educational Dashboards

Learning analytics aim to exploit the potential of the large amount of data describing interaction data, personal data, and information generated from online learning environments [Jiv18]. It can also bridge the gap between learning science and data analytics[Sie12]. Learning dashboards are a typical intervention method for learning analytics, which are a visualization tool to help instructors and learners make decision about the learning process.

Previous studies [Jiv18] clustered educational dashboards into six different education levels based on the competence they aimed to effect in learners: metacognitive, cognitive, behavioural, emotional, self-regulation, and tool usability. Metacognitive groups aspects related to learners' knowledge, beliefs, and reflection on their learning processes, strategies, and their effectiveness. The three aspects included here are understanding of the information displayed on the dashboard, agreement with this information, and the impact the dashboard has on learners' awareness and reflection. Cognitive level contains aspects that evaluate learners' understanding and knowledge regarding the studied material and was used through their performance and the quality of their learning outcomes. The impact of the dashboard on the learners' engagement, online social behaviour, and help-seeking behaviour are grouped under the Behavioural level. This level also included learners' usage of the learning environment and the dashboard itself. Emotional level consists of two aspects: impact on motivation and impact on affect. Finally, the tool usability level groups aspects related to the tool acceptance, case of use as well as learners' satisfaction.

Standard graphs used by prior work [Agu09; Gau08; Gol96; Har99; Har04; Maz03; Maz07] include scatter plot, bar indicators, bar and stacked bar charts, line graphs, pie charts, and heat maps. These student graphs provide a set of visualizations showing different kind of students' information or were used as a supporting graph in complex visualization. The works that provide visualizations with multiple categories of information usually have the goal of giving instructors a overall view of their students' learning experience. These systems provide an overview of the course and visualize students individually. Hardy's [Har04] e-learning tracking visualizations used node graphs to show a timeline of students' activities and allowed filtering by time and any subset of students. These activities extracted Virtual Teaming Environment log files to provide when and what student did in the learning system.

Akintunde et al. [Aki21a] developed PEDI (Piazza Explorer Dashboard for Intervention) to analyze and visualize students forum activity on Piazza, a question and answer forum. They detected troubling engagement levels and informed instructors through visualization. In this dashboard, PEDI first displayed the average post frequency and post frequency per day for instructors to aware the number of students demonstrating help seeking behaviours. Then it provided the forum activity for instructors to observe the number of students for different activities, i.e. not active, post followup, reply to followup, answer, and question. In addition, their prior research found a negative correlation between the number of likes on assignment related posts and the class average performance [Aki]. Thus, the PEDI dashboard displays a tree map that reflects the number of likes of posts to instructors. Gitinabard et al. [Git22] also developed a visualization tool that was focused on monitoring team collaboration. They first asked instructors to select the course, time frame, and what aspects of student activities they would like to see. This application would show the overview of a team, such as how many posts each team made on the forum and differences of team member from each team, in order to inform instructor of inappropriate team collaboration. This dashboard also displayed team activities on Github and MyDigitalHand [Smi17], an online ticketing system for office hours.

While all these dashboards designed for instructors provide valuable information about different aspects of the student activities, most focus on the demographic characteristics of students' activities, for example, the number of actions they did on different platforms. In addition, the discussion forum plays an important role in online learning courses. Thus, unlike other dashboards, we designed a content-based dashboard based on semantic analysis of students' post text. In this case, our designed dashboard provides instructors with students learning status from analyzing forum post content.

2.4.1 Dashboard Evaluation

Data used in evaluating dashboards came in three main ways: self-reported by learner, tracked automatically generated data from the system, and assessment data such as grades. Many of previous dashboard evaluations were based on feedback surveys and interviews. For example, [Alj13] built an analytical application called Quiz My Class Understanding (QMCU) which was purposely developed to investigate the significance of learning analytic techniques in order to provide students with immediate detailed feedback. They evaluated case study reports which also reflect the dashboard increasing users' engagement.

A large portion of previous dashboard evaluation studies used trace logs, which were collected by the platform, for assessing whether there were any changes in learners' use of resources. [Beh16] designed three learning analytic visualizations where each showed particular information about an aspect of students' participation in online discussion in a blended course. They evaluated the visualization based on linguistic features extracted from the input text, e.g., word concreteness, and text simplicity.

2.4.2 Instructor Interviews

"We, as educational technologists, often create innovations that are not adopted, despite a plethora of effective educational technology innovations" [Arc99; Tat08] and an increasing willingness and preparedness by educational institutions to implement technology [Joh06]. Diffusion of innovation theory predicts that we can increase the adoption rate of our innovations if instructors perceive our innovations solve their challenges [Rog10]. Educational technologists seek long term classroom adoption, or integration, of educational technology. We want instructors to use technology in regular classroom practices to support learning and maintained use over time. To achieve wide-scale long-term adoption, diffusion of innovations literature stresses that we should understand the challenges instructors perceive in their classroom practices, design technologies that help instructors overcome these challenges, and communicate these benefits. Designers of educational technology must balance the needs of multiple stakeholders [Eas18], particularly those responsible for adopting technology in a classroom. Consequently, a sound starting point for promoting adoption of educational innovations is an empirically grounded framework describing instructor perceived challenges to inform design and communication. Lewis et al. [Lew19] interviewed 47 university instructors about their most recent changes on authentic project-based learning (APBL). They investigated the challenges instructors experienced in APBL itself, an effective innovative pedagogical approach, to inform innovations that can be adopted and integrated into existing APBL classrooms. Zheng et al. [Zhe16] reported an interview study of 14 MOOC instructors in which they used ground theory to uncover the complex processes, motivations, and challenges associated with teaching a MOOC. A key finding they observed was that we we should provide support through the whole instruction process.

2.4.3 Course Monitor

Stephens et al. [SM14] surveyed 92 MOOC instructors to investigate three questions. First, what information sources do MOOC instructors prefer to analyze students online behaviours? Second, which of these sources are most useful to instructors from three aspects: course preparation, course administration, and course postmortem? Third, how should these data sources be visualized in tools for instructor to use? Based on their survey results, they first found that understanding forum activity was of interest to 97% MOOC instructors. Second, they found that MOOC instructors do not think chat logs are a valuable data source for understanding students' online behaviour. Third, MOOC instructors would like to receive the same sources of information as instructors of smaller-scale distance learning courses, as most of the MOOC instructors surveyed come from in-person class backgrounds.

Prior studies also developed a tool to help monitor in-person classrooms. Ngoc et al [NA19] developed an application based on computer-vision technology to allow faculty to capture and make a summary of student behaviours in the classroom. The system recorded the entire session and identified when students paid attention based on their position with other students. Schuck et al. [Sch16] designing a mobile application to monitor and improve classroom behaviour for children with Attention Deficit/Hyperactivity Disorder (ADHD). They examined 12 5th-grade students with ADHD. Their application prompted students every 30 minutes through a step-by-step self-evaluation process considering the last 30 minutes of classroom behaviour. At the same time, teachers also recorded behavioral observations for each student on another device. They found out that this application improved student self-awareness, self-regulation, and informed classroom behavior management techniques for students with ADHD.

Most of the prior dashboard studies focused on displaying the demographics and students' learning statuses, such as grades, assignment attempts, etc. Previous survey results show that discussion forums were the most frequently preferred source of information for MOOC instructors. Similar to previous studies on aspect of educational dashboard design [Agu09; Gau08; Gol96; Har99; Har04; Maz03; Maz07; Git22; Aki21a], we extracted student data from student activities and different from them, our designed dashboard focused on students' discussion forum activities from the aspect of posts' content. On the other hand, previous studies monitor students forum actives from the aspect of quantity, such as how many questions/answers student post in the forum and detect whether the student is active in the forum [Aki21a; Git22]. Unlike them, we focused on monitoring student post content to detect the hot topics students discussed. Previous studies [SM14; NA19; Sch16] have shown that monitoring students in both online and in-person learning environments and monitoring students' activities could help instructors to understand students learning status and make interventions to support students' learning. Thus, based on the fact that the lack of tools monitoring students' social learning and help seeking behaviours from the content of forum, we designed a content-based dashboard based upon students' forum activities in order to help instructors monitor students' online learning statuses. In addition, we designed interview questions based on prior dashboard evaluation studies to investigate the usefulness and the usability of the dashboard.

Chapter 3

DATASET

3.1 Data

To address our research questions, we analyzed data from "Big Data in Education" MOOC offered by The Teacher's College at Columbia University and hosted on the Coursera (BDE 2013) and EdX (BDE 2015) platforms in 2013 and 2015, respectively.

The "Big Data in Education"(BDE) MOOC was designed to increase interest in educational communities to use data mining methods to answer educational research questions and improve interventions in educational systems. This was an eight-week course that included materials for a graduate-level course on educational data mining and big data analysis in education. The course was comprised of weekly lectures in the form of videos, individual assignments, and quizzes which contributed to students' final grades. The students were required to answer numeric or multiple choice mastery questions in the weekly assignment by applying data mining techniques from the lectures. In the BDE MOOC, students and instructors participated in the forum every week. Forum participation was not required for students and did not count towards their final grades. For this study, we considered the category of forum posts as their forum actions.

The course had a total enrollment of over 48,000 students, but a much smaller number of active participants. 13,314 students watched at least one video while 1,242 watched all of them. A total of 1,380 students completed at least one assignment, and 778 made at least one post or comment in the forum. Of those students who made with posts, 426 completed at least one class assignment. A total of 638 students completed the online course and received a certificate (meaning that some were able to earn a certificate without participating in forums at all). In order to achieve 'Completion', students had to receive an overall grade average of 70% or above. In the 2013 class, 778 students made at least one post or comment on the forum with a total of 603 discussion threads consisting of 4259 posts in total. In 2015, 519 students produced 625 discussion threads with a total of 2056 posts. However, in contrast to Coursera, which runs all communication through a single channel, the EdX platform includes a private chat feature which is used for individual discussions. We were not able to obtain access to this chat data for our analysis.

3.2 Coding Schema

In order to provide instructors insight of students' learning status, we built our coding schema based upon qualitative observations from the entire dataset as described in the Chapter 5.4.1. Our coding schema consisted of 12 classes representing different types of cognitive engagement during forum discussions. Listed below is a concise explanation for post types and their abbreviations in order of frequency.

- *On topic question (Q)*: Questions related to course topics or homework completion.
- *On topic statement (O)*: Answers to questions or opinions on Q questions.
- *Question about technical issues (BQ)*: Technical questions about coding or software issues.
- *Technical issues statement (B)*: Answers to questions or opinions on BQ questions.
- *Course logistic questions(CQ)*: Questions about course logistics.
- *Course logistic (C)*: Answers to questions or opinions on CQ questions.
- *Announcement (A)*: Announcements to the class.
- *Additional materials (M)*: Provisions of supplemental materials or videos for students.
- *Social connection (S)*: Students' self-introduction to the class or attempts to set up discussion groups.
- *Politeness (P)*: Expressions of politeness or expressions of gratitude to classmates or instructors.
- *Off topic (T)*: Irrelevant posts or broad complaints about the instructor or the course.
- *Other language(X)*: Posts not in English.

Figure 6.1 shows the hierarchical structure of our coding schema. Previous studies investigated the significance of distinguishing content as related and non-related posts and questions versus non-questions. [Cui15; Wis16; Wis18; Kul18]. They found that students who contributed to related threads received higher final grades than those who did not. Thus, in order to examine the relationship between post types and student's course performance, we first determined whether a post is relevant to the course content or not. Then for these on-topic posts, we explored whether it was a question or a statement to utilize information for a future Question-Answering System. For questions or statements, we determined what topic the student was discussing, i.e. Technique/Coding, Logistic, or Course concept. In addition, individuals may show politeness and make a statement at the same time. In this case, we consider this a statement type rather than simple 'politeness'.

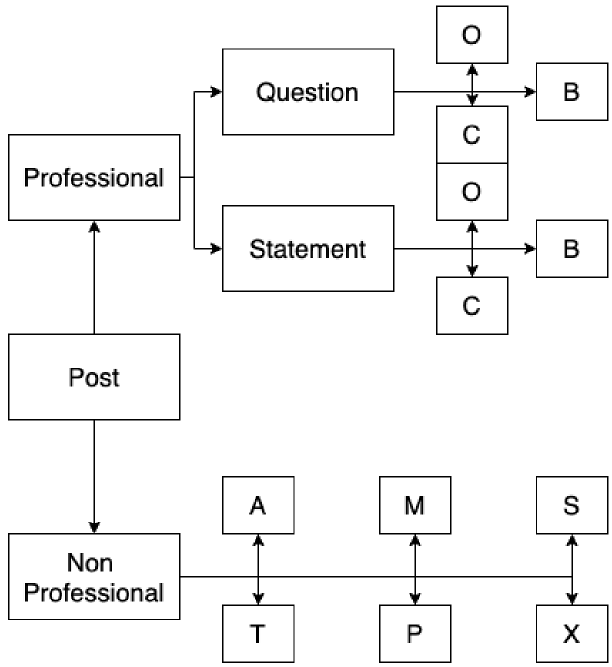

Figure 3.1 Coding schema

Three experienced researchers annotated the two datasets using our coding schema. Two researchers annotated BDE2013 and two annotated BDE2015 with the lead author annotating both. We calculated inter-rater reliability using Cohen's kappa [Ber88]. Kappa represents the discrepancy between the observed probability of success and the probability of success under random cases.. Independence implies that the pair of raters agree about as often as two pairs of people who make random decisions. The range for kappa value is from 0 to 1. In general, an agreement at 0.8 is considered high annotation quality, while sometimes a coefficient above 0.7 is good enough for text analysis [Ind10; Vie05]. To measure the inter-rater reliability, we randomly selected a 20% data sample from both datasets, and had three researchers independently label them. We obtained a 0.81 kappa for BDE2013, and 0.71 for BDE2015 which means that for both datasets, we maintain a good agreement when coding posts.

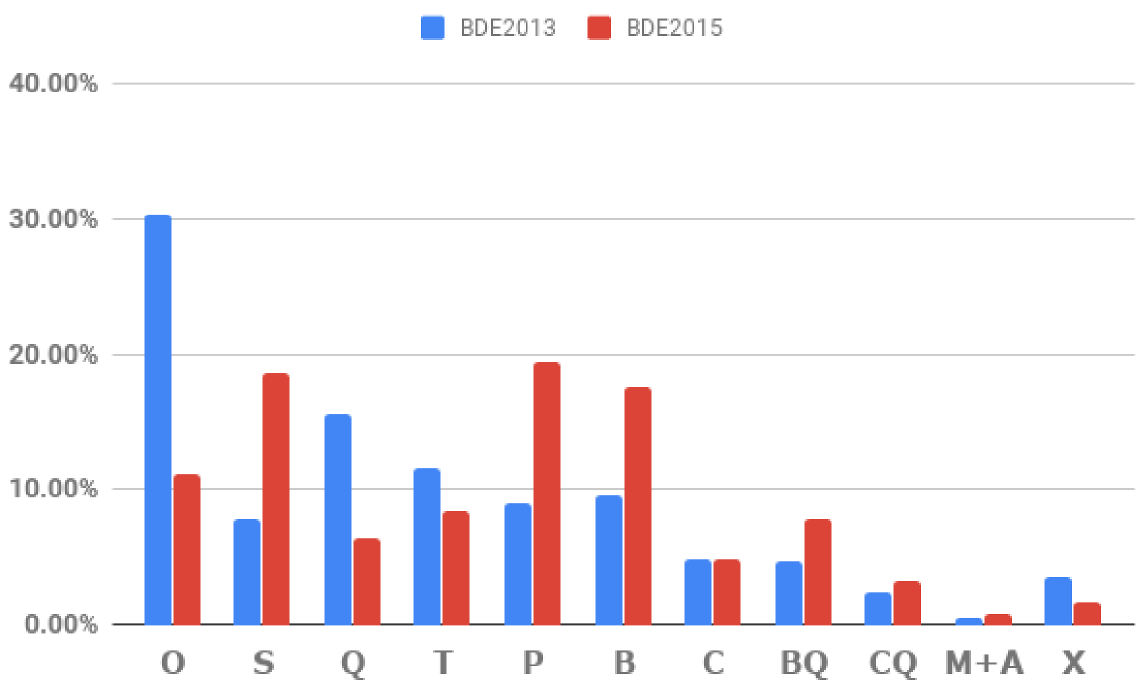

Figure 3.2 Class distributions for each course

Chapter 4

RQ1.1 - How do students build their communication over time based on their forum reply relationship?

Massive Open Online Courses (MOOCs) are designed on the assumption that good students will help poor students thus offloading the individual support tasks from the instructor to the class. However prior research has shown that this is not always true. Students in MOOCs tend to form distinct sub-communities and their grades are closely correlated with those of their closest peers. That work, however, was only based on analyzing the final social network in a MOOC. In this paper, we study the evolution of these co-performing clusters over time. We explore a longitudinal approach to detect how students form their social connections on the discussion forum and we show that students form close coequal communities early in the course and maintain them over the duration of the course.

4.1 Introduction

One promise of Massive Open Online Courses (MOOCs) is that we can provide high-quality educational content to students around the world at relatively low cost. The broad goal of MOOCs is to scale instruction by allowing expert instructors to provide guidance to hundreds or even thousands of students at a time. Such large-production education has the potential to be revolutionary both for individual students and for educational systems. The current generation of MOOCs are designed to achieve this scaling by outsourcing much of the individual support tasks to students. That is, rather than capping enrollment to ensure that the instructor and TAs can support every students' needs, MOOCs provide online forums that encourage students to share common questions and to provide collaborative guidance or to benefit from each others' interactions with the limited support staff. Thus it is tacitly assumed that students will have common issues and that good students will help poor students with course content, assignments, logistics, and other issues. The role of instructors and TAs is then often to *curate* help rather than *authoring* it.

In a prior study Brown et al. examined the formation of communities in a large scale MOOC on Big Data in Education [Bro15b]. They extracted the social network structure of the course from the online forum and analyzed the connections between students. Contrary to the implicit assumption described above, they found that the social connections were not evenly distributed. Nor did they find that the lower-performing students made persistent connections with their higher-performing peers. Instead they found that the students formed distinct sub-communities and that their performance in the course was strongly correlated with that of their closest neighbors. In followup work, Brown et al. also found that these communities were not aligned with students shared backgrounds nor were they apparently driven by shared course goals [Bro15a]. They also found that these results were stable even after the instructional staff and other highly-connected or *hub* students were factored out. Thus the authors concluded that the pattern of students' social relationships can be used to predict their performance and that interventions which target those social relationships might help students to improve either by selecting good peers or by flagging isolated and poorly-performing groups for individual attention.

That work, however, was limited by the fact that it only used the *final* social network from the course. Thus when evaluating students' performance the authors included all posts and social interactions that had developed over the duration of the course. In order to provide useful guidance during the course and to provide reliable information to instructors, we must show that it is possible to detect these relationships based upon partially-formed networks. In general most students' help-seeking patterns change over the duration of the course. Students often drop out of courses, particularly MOOCs, or taper off their involvement as they lose interest. Students also face difficulties in courses which prompts them to scale up their communication as the course becomes more challenging. It may be the case that the structure of the network will change radically over the course of the class and that any early detection model or instructor dashboard will be erratic, invalid, or simply out of dated.

In this work, we expand upon the prior work of Brown et al. by examining the growth of the students' social relationships over time. To that end we segmented the forum data by time and performed a sequential analysis of the evolving social network. Our goal in this work will be to address the following questions: First, are students' social groups stable over time? And if so, how early in the course do these observed grade relationships hold? Second, can we use partial social networks to help inform instructors and students in MOOCs? If the answer to these questions is true then it may be possible to develop effective social intervention systems that could use students' posting behaviors to flag students that need attention, or to key off strategic advice on where or how often to post questions. Section 2 provides some background on social network analysis in education. Section 3 describes the dataset we use in our work. In Sections 4 and 5 we present our analysis and results. And finally in section 6 we present out conclusions and discuss our future work.

4.2 Background

4.2.1 MOOCs, Forums, Students Performance

According to Seaton et al. most of the time students spend on MOOCs is spent viewing the lecture videos, completing mastery assignments, and reading the discussion forums [Sea14]. Very little time is spent on external or 'off-platform' activities. Thus the discussion forums provide a rich and useful window into the students' primary course activities. Stahl et al. [Sta06] illustrated how students collaborate to construct knowledge through this interaction. They argued that students' forum activities are not only beneficial for the individual discussants but also serve to structure the class as a whole. Each student's activity level varies as does their impact on the course. Huang et al. for example, specifically investigated the behavior of high-volume posters in 44 MOOC-related forums. These 'super-posters' tend to enroll in more courses and generally perform better on average [Hua14]. Moreover, by actively engaging in many conversations, they add to the overall volume of the course discussion and they tend to leave fewer questions unanswered in the forums. They also found that, despite their high output, these super-posters did not act to suppress the activity of other less-active users. Rienties et al. [Rie14] examined the way in which students structure their social interactions online. They found that allowing students to self-select collaborators in a MOOC is more conducive to learning than random assignment of partners. In another study, Van Dijk et al[VD01]. found that simple peer instruction is significantly less effective in the absence of a group discussion step, thus reinforcing the importance of a shared class forum.

Prior researchers have also examined the general social and behavioral dynamics of the student forums. Boroujeni et al. examined the relationship between students' temporal patterns, discussion content and social structures emerging from the forums [SB17]. They found that for eight week-long MOOCs, the pace of students' posts remained high during the first 3 weeks and then tapered down gradually until the class ended. They also found that this pattern was affected by the assignment dates and other deadlines as well as the overall volume of the posts in each thread. Furthermore, they tracked the network attributes over time by using one-week network slices based upon a sliding window. The slice for each day of the course (d¿6) was built from forum activities during the peceeding 7 days ([d-6, d]). For each network slice, the attributes included node counts, edge counts, average degree, density, etc. They found that, with the exception of density, the attributes decreased over time. Density, by contrast, increased sharply at the end of course. Zhu et al. explored a longitudinal approach to combine student engagement, performance, and social connections by applying exponential random graph models [Zhu16]. They analyzed the relationship between the social networks on a week-by-week basis and they found that students' individual assignment scores were all positively related to being more active in social network.

Rosé et al. [Ros14] examined students' evolving social interactions in MOOCs using a Mixed-Membership Stochastic Block model which seeks to detect partially overlapping communities. Their specific focus in the analysis was on identifying the students who were most likely to drop out. They found that it was possible to predict whether or not a student would drop out based upon their membership in a community. Students who actively participated in the forums early on in the course were less likely to drop out later on. Moreover, they found that one specific sub-community was much more prone to dropout than the remainder of the class. This suggests that the forum communities do align by stability and thus that social relationships can reflect the students' relative level of motivation as well as their overall experience in the course. This is akin to the 'emotional contagion' model used in the Facebook mood manipulation study by Kramer, Guillroy, and Hancock [Kra14].

Dawson et al. [Daw10] elaborated the use of social networks to provide guidance. They provided feedback to students and instructors based upon the students' *ego-social* network (i.e. their neighborhood). They explored differences in the network composition for low- and high-performing students to identify patterns of behaviours which may influence the students' learning. They found that the ego-social networks of low- and high-performing students had significant differences, it was possible to identify different types of students based upon their ego-network. They also found that the instructors were equally likely to show up in high-performing students' local networks as in those of the low-performing students. Their results indicated that instructors could adjust their teaching methods based upon this network structure.

4.2.2 Communities

A particular area of concern is how students connect within sub-communities and with the instructor. Insa et al. showed that in a traditional course (containing both face-to-face lectures and lab sessions), the student's seating position can affect their final grade [Ins16]. They suggested that physical proximity to the instructor increased performance. According to Golder et al., an analysis of students' Facebook messages showed that the students will message one another more often during weekday afternoons than over the weekend [Gol07]. This produced a distinct temporal pattern in their communication and community structure.

The motivation for any student to join a MOOC can vary widely. This can in turn create several distinct classes of participants with their own unique behaviors. Anderson et al., for example, argued that MOOC participants can be partitioned into 5 distinct categories based on the number of lectures that they watched and on the assignments that they submitted: viewers, solvers, all-rounders, collectors, and bystanders [And14]. They also found that the more assignments a student completed and the more lectures that they viewed, the higher their final grade would be. Interestingly, while students who received a 'B' grade showed a small decrease in their homework submissions relative to 'A' students, the amount of time that those students spent watching lectures was substantially lower. In related work by Liu et al. however, the authors found that some of these behavioral differences were consistent with the students' cultural background which may affect not just their motivation but their expectations and habits [Liu16].

Other authors have examined the relationship between students' academic performance and their social network relationships. Eckles et al. used network analysis on survey data to identify at-risk students who were more likely to drop out [Eck12]. Kovanovic et al. analyzed how a student's relative centrality in their social network will affect their academic performance [Kov14]. They found that more central students were typically higher performers than their less-connected peers. And finally Zhang et al. constructed student social networks based upon the comments and replies that had been posted to the forum [Zha07]. By analyzing the relative in- and out-degree of the vertices, they were able to identify a small amount of users who answered a large proportion of the questions. This allowed them to find key students in the course.

4.2.3 Student Behaviours

In their analysis of student behaviors, Anderson et al. found that the number of students who watched lecture videos and finished assignments decreased over the duration of the course, suggesting that some students changed their minds about the class or simply changed their habits during it [And14]. Ye et al. performed a similar study, in which they examined a 10-week computer science MOOC [Ye14]. At the end of week 4, 60% of the students who had only watched lectures but had not participated in other ways had dropped out of the course, while only 20% of the students who had submitted assignments and completed quizzes along with viewing had done so.

Given that a large number of MOOC registrants in each course drop out [And14; Ye14], studying the causes of this dropout and preventing it is an important issue. Kloft et al. sought to predict dropout behaviors in a 12-week course based upon the students' click-stream data using a Support Vector Machine [Klo14]. They identified two peak dropout points, one during the first two weeks of the course, and the second at the end of weeks 11 and 12. Students were unlikely to drop in the middle of the course and thus if they made it through the early stages and the final crunch then they would likely complete. Halawa et al. used a specialized definition of drop out as a student being absent from the course for more than 1 month or if they viewed less than half of the lecture videos [Hal14b]. With this definition they found that the percentage of students absent from the course sharply decreased from 36.4% to 13.8% after week 3. Hoskins , by contrast, focused exclusively on self quizzes performance-based indicators and found that low performing students tended to drop out more than their higher-performing peers [Hos05].

4.3 Methods

We began our analysis by clustering the count of students' submissions for each assignment by date in order to understand when students completed their assignments and how the submission patterns might indicate their working habits. Unsurprisingly the assignment submissions peaked right before each due date with few if any late submissions. In order to make our analysis more consistent we broke the 8-week course into 2-week chunks and we split our analysis at weeks 2 (start), 4 (midterm), 6 (third quarter), and 8 (final). This decision was based upon the fact that students worked across weeks and on prior literature which pegged the two-week and four-week boundaries as crucial times for student dropout (e.g. [Klo14; Ye14]).

This partitioning yielded four distinct datasets representing the cumulative forum discussion up to that point in the class. We extracted a social network from each of these datasets using the same approach applied by Brown et al. [Bro15b; Bro15a]. In this approach we generated a raw social network for the course where each node represents a single participant (student, TA, or instructor). We then labeled the student nodes with their cumulative performance up to the specified time step. Thus, the week 2 dataset was labelled using their cumulative performance up to the end of week 2. The Coursera forums operate as standard threaded forums. Users have the ability to start new threads by making an initial post. They can also add posts to the end of an existing thread or add a specific reply below a given post.

In order to build social network from discussion forum, we considered participants as nodes and their communications as edges. More specifically, for each comment in a thread, we added a directed arc from the author's node to a node representing the author of each comment that precedes it in the thread, with the exception of self-loops. So all of the contributors to a thread will be connected with one-another including the originator. This approach is based upon the assumption that students read the thread *before* contributing to it and that a contribution to a discussion represents a contribution to the whole conversation. The average length of each thread in our dataset was seven posts. Thus we treat each reply as evidence of an *implicit social connection* between the individual author and their conversational peers. Such implicit social relationships have been explored in the context of recommender systems to detect strong communities of researchers [Cho14]. The resulting networks form a multigraph with each edge representing a single communicative act. As our goal is to focus on social relationships we then modified this graph by eliminating all isolated nodes, and by collapsing the parallel edges to produce a weighted undirected simple graph representing connections between students.

In addition to analyzing the general connections between students, we also sought to analyze the impact of the instructional staff and the active hub students on their social structure. We therefore generated three different graphs for each of the datasets: *ALL* which is the complete graph with all non-isolated nodes; *Student*, which eliminates the instructional staff; and *No Hub*, which removes both the instructional staff and the highly active 'hub' students. Since MOOCs are an at-will course students often drop out and we cannot always distinguish intentional dropouts from unintentional failure. In one typical dataset, for example, more than 80% of the students received a grade of 0 [And14]. Therefore we also constructed graphs for students with and without students who received a grade of 0.

4.3.1 Best-Friend Regression

Fire et al. modeled students' social interactions for grade prediction in a traditional classroom [Fir12]. They found that in traditional classes the student's grades can be closely correlated with those of their closest neighbor or "best friend". That research was based upon self-reported relationship data, but Brown et al. were able to show that it also applied in an online context [Bro15a]. In that analysis they used the weighted network to identify each students' "Best Friend" (BF) or closest peer by connections. They then showed that the same result held for this network structure as well.

4.3.2 Community Detection

We applied the Girvan-Newman algorithm to find social clusters within our graph. In order to identify the ideal number of clusters we used the "natural cluster number" approach described in [Bro15b]. That approach is based upon the modularity score of candidate clusters. Given a graph that has been clustered into sub-communities, the modularity of the graph is measured by the ratio of intra-cluster to inter-cluster connections. That is, how strongly are individual students associated with their cluster associates relative to the rest of the class. Graphs with high modularity have very strong within-cluster connections and relatively sparse connections across the groups. As the graphs are partitioned into smaller and smaller communities the modularity score will grow rapidly until we reach an inflection point or a point of diminishing returns at which point each additional sub-cluster makes little difference to or even reduces the modularity score. In the natural cluster approach we iteratively cluster the graph into higher numbers of communities and plot the modularity score over number of clusters. We then examine this curve to find the inflection point and use that value. This is an exploratory approach similar to exploratory Principal Components Analysis.

4.3.3 MOOCs, Forums, Students Performance

In MOOCs, the class forum is typically the only official way for students to communicate with the instructors and with each other. Thus, their activities on the forum represent a mostly-complete record of their communicative actions and it represents the best record of their questions and interests. So the dynamic of student forum activities represent their real-time learning status. In order to investigate the dynamics of the students' forum activities and their relationship with the students' social networks, we extracted number of posts and comments, number of forum users (who wrote a posts) and the number of threads added on a biweekly basis. We then analyzed the numbers in each two-week pair to find the scale of the social network in each case. We also explored how the social aspects of discussion forum changed over time, by calculating density, degree, average path, diameter and other basic metrics. These network attributes represent the evolving network structure. Futhermore, we compared the scale of the dynamic networks and the network structures to determine when the social networks stabilized.

And finally we analyzed the average number of changes for each student neighbors to learn how students selected their communities biweekly.

4.4 Results & Discussion

Table 4.1 shows the order (number of nodes) and size (number of edges) of the graphs that we obtained at each cutoff point. As the table illustrates while the graphs grew monotonically in order and size over the duration of the course, most of the connections between the students were already established by the end of week two. That is, the basic network structure, if not its weight, was set early on.

Table 4.1 Graph order and size for each cutoff.

	Nodes	Edges	Comments
Week2	645	14,050	2,472
Week4	693	15,142	3,231
Week6	725	16,346	3,833
Week8	754	17,004	4,260

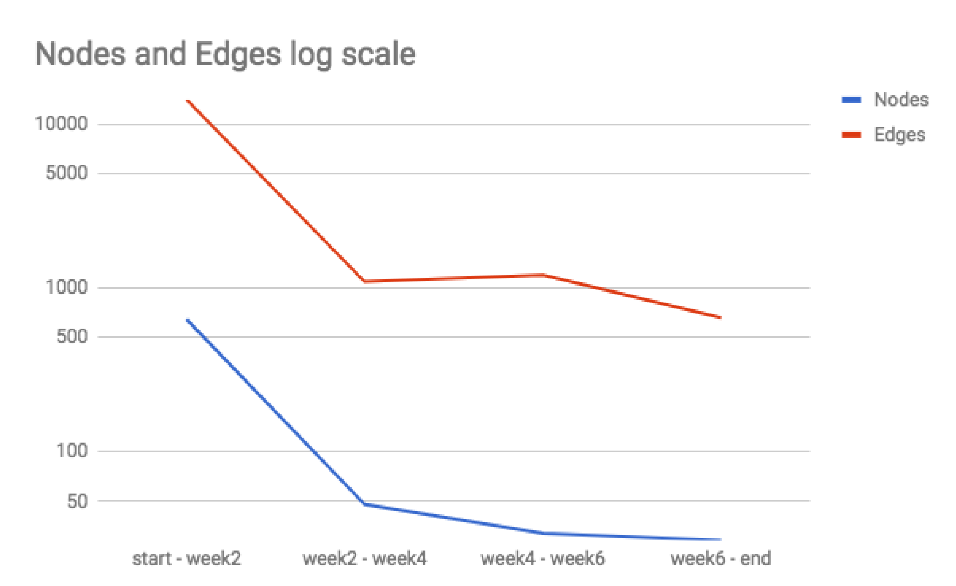

Figure 4.1 Newly participating students and connections every two weeks

At the end of the course, there were a total of 55,179 registered users, yet the final course graph contained only 754 participants, 751 of whom were students with 1 instructor and 2 Teaching Assistants. Additionally, 304 of the 751 students obtained a zero grade at the end of the course while 447 received non-zero grades. Some of the forum participants did not complete any assignments but still chose to discuss the course topics with others. By the same token, some of the students who completed work in this course did not participate in the forum at all. There were 1,381 students who received a non-zero final grade 934 of whom did not post in the forum, while 304 zero final grade students did. It is conceivable that when the students met with problems, they chose to ask questions online, but participation in the course forum was not a necessary condition for completion.

We observed that most of the students who used the forum had made at least one post by the end of week two. Only a few students began communicating after that time. Figure 4.1 shows the number of new participants and new connections added into the social network every two weeks. We applied log scale for the y-axis to make the chart more readable. As these results illustrate almost all students and instructors had established their connections in this course by the end of week two and only few of new connections were made after that time. Additionally, the total number of posts/comments made was 4,260, 2,472 of them (or 58%) had been made at the end of week 2. Even if we discount the 200 or so 'introduce yourself' comments, it still shows that most of the posting activity happened at the beginning of the course. One potential explanation for this is that the students, particularly those who did not plan to obtain a certificate, did most of their work early and subsequently lost interest. Or that some of the students worked in binges and did not fit the schedule over time. An ongoing analysis of the forum content has shown that a number of the posts are also about early issues such as course logistics and software, problems which may be less relevant later on. Irrespective of the cause it indicates that the social structure is well established early enough that information based upon it can be used to advise students before it is too late.

4.4.1 Best-Friend Regression

As part of our analysis we also replicated the Best-Friend comparison used by Brown et al. Here we identified each student's closest neighbor in the course, ignoring teaching staff, and we calculated a direct correlation between their grades and those of their best friends. Because the data was non-normal we used Spearman's Rank Correlation Coefficient (ρ) a nonparametric measure of association [Sed14; Dal02]. Our results are shown in Table 4.2. Because week 8 is the last week of the course, the intermediate grade is the final grade.

From Table 4.2, the students' grade and their best friends' grades, both final or intermediate grades for each bi-week, were strongly correlated, ρ was high, and significant $p < 0.001$. However, the correlation ws affected by the clusters of 0 grade students. After removing these students, the correlations did not hold at a statistically-significant level until the middle of the course. After week four, we found a moderate correlation, $\rho = 0.295$, $\rho = 0.25$, $\rho = 0.29$ and $p < 0.001$. Thus, the relationship between students' grades and those of their best friends were consistent

Table 4.2 Correlation and p-values for Best Friends analysis.

	intermediate grade		final grade	
	ρ	p	ρ	p
Week2	0.25	< 0.001	0.27	< 0.001
Week2_non0	0.086	0.12	0.093	0.08
Week4	0.313	< 0.001	0.339	< 0.001
Week4_non0	0.145	0.005	0.158	0.002
Week6	0.42	< 0.001	0.437	< 0.001
Week6_non0	0.25	< 0.001	0.295	< 0.001
Week8	NA	NA	0.44	< 0.001
Week8_non0	NA	NA	0.29	< 0.001

from the traditional face-to-face class to MOOC but not immediately. Our results show that MOOC students except those who did not submit any assignments performed similarly to their closest peers.

4.4.2 Community Detection

Table 4.3 Modularity and number of clusters for each graph with intermediate grade

Graph Type	Week2	Week4	Week6
All	112	177	200
Modularity	0.346	0.327	0.276
All_non0	56	100	121
Modularity	0.276	0.195	0.122
Students	119	129	172
Modularity	0.414	0.419	0.393
Students_non0	63	97	125
Modularity	0.436	0.346	0.266
Nonhub	63	67	69
Modularity	0.590	0.590	0.553
Nonhub_non0	43	41	55
Modularity	0.613	0.490	0.396

Figure 4.2 provides an example of the modularity curves both with and without zero-score students. We selected natural cluster numbers from these plots where has the largest modularity score. Table 4.3 shows the selected number of natural clusters based on each week's intermediate

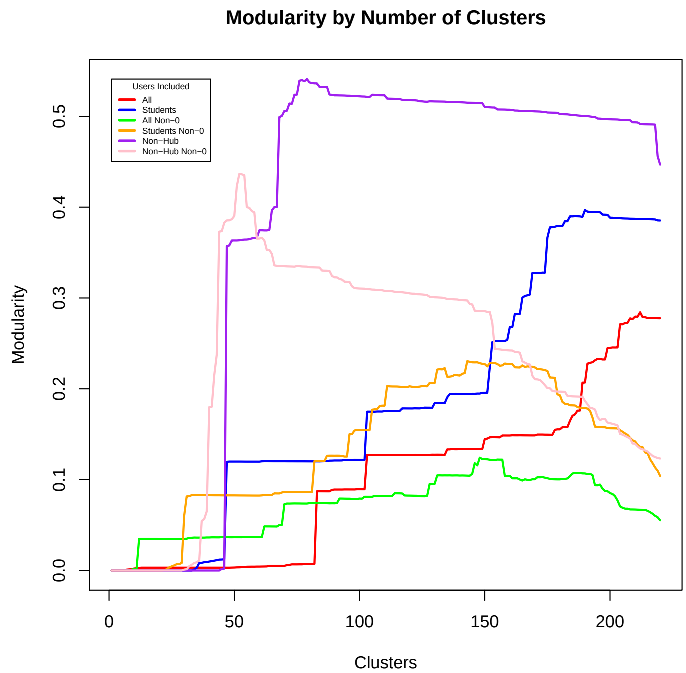

Figure 4.2 Modularity by Number of Clusters for Week 8

grade and table 4.4 shows the number of clusters based upon the final grade. From table 4.3 and table4.4, we found the maximum modularity score for clusters decreases over time. As the modularity score is designed to measure the cleanliness of dividing the network into clusters, these results indicate that the connections between the individual students are more sparse while the connections between the clusters of student are more dense.

Interestingly, the curves for the ALL and Hub Student graph are extremely similar, which indicates that hub-students were those who kept a close connection with instructor and TAs. As we anticipated, the non-zero students are the largest group of the students. The social network graph shows that many of the zero score students were only connected with other zero score students which supports our argument that poor performing students are likely to connect with others at their level.

In order to assess the cluster stability we also calculated student-centric cluster similarity metrics for the graphs. Tables 4.5 and 4.6 show the average number of neighbors that each student loses, gains, or retains in their cluster from week to week. That is, it shows how many former friends are now in a different group, how many new friends are added, and how many stay the same. These figures are shown for weeks 2-4, 4-6, and 6-8 for the all but the no-hub graphs. We excluded the no-hub graphs from our analysis because the definition of a hub student changed from week to week thus inflating the churn. We also generated the metrics for the social networks based upon the final grades and the weekly cumulative grades. As the tables illustrate the clusters lost members in each week with the losses being highest in the

Table 4.4 Modularity and number of clusters for each graph with final grade

Graph Type	Week2	Week4	Week6	Week8
All	112	177	200	212
Modularity	0.346	0.327	0.276	0.284
All_non0	56	135	149	173
Modularity	0.257	0.202	0.161	0.103
Students	119	129	172	184
Modularity	0.414	0.419	0.393	0.390
Students_non0	63	109	130	169
Modularity	0.439	0.351	0.304	0.224
Nonhub	63	67	69	79
Modularity	0.590	0.580	0.553	0.541
Nonhub_non0	43	45	49	52
Modularity	0.570	0.478	0.407	0.437

jump from week 2 to week 4, when the network is still growing faster. In the later weeks the losses were smaller, particularly in weeks 4-6. And, for all but the All_NonZero graph, the students gained few new neighbors with most of the neighbors being retained. As we discussed above the number of clusters increased as the course went on. As these tables indicate the new clusters were generally subsets of the prior clusters and did not present a remix of the prior neighborhoods. The lone exception was the All_NonZero graph which had substantial gains in weeks 4-6 and 6-8. This suggests that the lurkers and other non-certification-seeking students are an important factor in the stability of the social networks and thus that discarding them has a notable effect. However more analysis is required to understand just how they engender this stability and just how widely distributed they are in the clusters.

4.4.3 Student Performance & Motivation

According to the social network graph, students clustered into different clusters based on their connections and their performance. In order to examine the grade distribution of each cluster, we applied the Kruskal-Walls(KW) test to evaluate the correlation between clusters and performance. The KW test is a non-parametric rank-based similar to the common Analysis of Variance [Kru52]. The result for each graph shown in table 4.7 - 4.8 while the 'F' column value is Chi_square. We can see that for non 0 score students, their performance was highly related with their clustered friends, but when all students are included, the relationship becomes weak. This result supports our hypothesis that students will connect with similar performers, instead of helping poor performing students or learning from good ones [Wan15b].

Table 4.9 is representative of the evolution of the forum attributes over the 2 week intervals. The overall number of posts, threads, and users increase over time. From the table, we can see that the increase in the number of posts and threads is stable from course start to end. By the end of week 2, 59.4% of the posts had been added to the data and 57.8% of the threads

Table 4.5 Average Dynamic Cluster Changes with final grades

lost: average number of lost neighbors
gain: average number of new neighbors
overlap: average number of the same neighbors

finalgrade	all			all_non0		
week	lost∗	gain∗	overlap∗	lost	gain	overlap
2-4	11.7	1.75	29.6	8.46	2.63	9.94
4-6	1.75	1.75	28	2.53	17.9	9.3
6-8	1.9	3.27	26.74	9.86	36.3	16.9
	students			students_non0		
week	lost	gain	overlap	lost	gain	overlap
2-4	2.05	9.47	30.7	19.3	2.61	11.4
4-6	9.7	1.55	28.77	3.8	2.96	9.42
6-8	1.64	2.72	27.7	2.72	8.94	9.34

came into the course forum. However, consider the number of users, 85.6% of the total forum users showed up by week 2. So, by one quarter of the way through the class most of users had already showed up in the forum but fewer than 60% of posts and thread had been initiated. Table 4.10 shows that the values of the network attributes don't have clear changes which may indicate that root social network structure doesn't change after week 2. Thus, the dynamics of the forum attributes are consistent with our findings for the best friends and community analysis over time, that the student forum social network structure will hold as soon as week 2 and will then become stable with the small communities and best friends only getting stronger.

4.5 Conclusion

Our goal in this study was to address the potential utility of social network information to guide students and instructors in MOOCs. As prior work has shown students' final social network structures, particularly their closest neighbors or "best friends" and their sub-clusters, can be analyzed to predict their performance. However, in order to provide meaningful guidance, or to help students and instructors improve their performance before it is too late, it is necessary to show that we can extract useful information from partially-formed social networks. In this study, we have done just that by showing that the structure of the students' social networks can be analyzed to predict their performance even by the second week in the course.

Consistent with the prior literature we found that students are most closely associated with similarly-performing peers and it is possible to predict students' performance based upon their closest neighbors in the graph. Thus in contrast to existing assumptions good students are not necessarily connecting closely with poorer performers or spreading their help evenly across the class. These results hold even if we remove the instructional staff, hub students, and zero-grade

Table 4.6 Average Dynamic Cluster Changes with intermediate grades

intergrade	all			all_non0		
week	lost	gain	overlap	lost	gain	overlap
2-4	11.7	1.75	29.6	14.6	20.2	17.1
4-6	1.75	1.75	28	6.87	50	28.8
6-8	1.9	3.27	26.74	41.7	15.3	37.7
	students			students_non0		
week	lost	gain	overlap	lost	gain	overlap
2-4	2.05	9.47	30.7	20.5	3.2	11.2
4-6	9.7	1.55	28.77	3.28	17.5	10.3
6-8	1.64	2.72	27.7	17.3	8.2	10

Table 4.7 KW test with intermediate grade

Graph Type	Week2		Week4		Week6	
	F	P	F	P	F	P
all	207	< 0.001	270	< 0.001	315	< 0.001
all_non0	74	0.04	133	0.07	129	0.25
students	218	< 0.001	228	< 0.001	285	< 0.001
students_non0	55	0.69	118	0.06	142	0.12
nonhub	134	< 0.001	171	< 0.001	182	< 0.001
nonhub_non0	53	0.1	47	0.18	90	0.001

students from the course.

Thus it is theoretically possible to analyze this data to support instructors by flagging isolated students or poorly-performing sub-communities that are in need of help. It might also be useful for direct support by allowing us to provide guidance to students who may not be seeking help from the right places. By identifying students who are not isolated but who are not necessarily getting help from good peers we may be able to intervene and to not only improve their individual standing but also the structure of the course as a whole. These results also suggest that we should consider mechanisms to encourage more distributed feedback such as explicit awards for peer tutoring.

Interestingly we found that students' social behaviours are consistent because, while students continue to contribute to the course over time the social structure of the course is established relatively early. More than half of the forum posts are made in the first two weeks of class. And few students begin to participate on the forum after that point. Thus it is not the case that we have a dynamic graph which can be analyzed differently at each stage. Rather it appears that the basic structure of the social relationships are fixed early and then only grow stronger over time. While more analysis is required to determine why this occurs it suggests that students'

Table 4.8 KW test with final grade

Graph Type	Week2		Week4		Week6	
	F	P	F	P	F	P
all	210	< 0.001	273	< 0.001	319	< 0.001
all_non0	70	0.19	154	0.1	168	0.12
students	223	< 0.001	239	< 0.001	293	< 0.001
students_non0	80	0.06	127	0.1	164	0.01
nonhub	145	< 0.001	179	< 0.001	190	< 0.001
nonhub_non0	44	0.2	58	0.06	67	0.03

Table 4.9 Forum attributes over time

Attribute	Week2	Week4	Week6	Week8
Posts	2514	3231	3833	4233
Users	659	707	742	770
Threads	345	460	545	597

initial impressions or choices have a strong impact on their performance and that interventions which are designed to change those habits may be beneficial. One avenue of research that we are currently pursuing is to analyze the content of the individual posts. If we can detect a change in the nature or structure of the content or of the topics being considered it might help to explain why the students' progress appears to taper off so dramatically. At the same time we plan to experiment with evaluating metrics of this type for blended courses to see if similar dynamic results hold in face-to-face contexts.

Furthermore, our results indicate that a social network analysis of the discussion forum data brings an unprecedented opportunity for instructors to visualize students' social structures and to form learning networks which allow them to make changes to their teaching plans over time. For non zero grade students, the correlation between students' grades and their best friends' grades is not reliable during the first 4 weeks of the course. However given the success of the network information, much could be done to detect at-risk students. This will ensure that the instructors might be able to identify these students in the first instance. Additionally, for low-performance students, their real-time ego-networks may explain why they are performing so poorly. Thus, low-performing students know know that if they wish to improve their grades, they will seek more connections with, or advice from, higher-performers.

Table 4.10 Network attributes over time

Attribute	Week2	Week4	Week6	Week8
Degree	21.783	21.850	22.546	22.552
Density	0.034	0.032	0.031	0.030
Avg_path	2.607	2.535	2.492	2.490
Diameter	7	7	7	7
Connected component	82	88	89	98

Chapter 5

RQ1.2 - How do we understand students forum post by categorizing the post content?

MOOC discussion forums contain a rich record of students' contributions, misconceptions, and conceptual change. However, it is often difficult to find structure in this semi-structured discourse, to track students' changing views, and to guide student learning over time. In this work, we developed automatic classifiers for student comments in MOOC discussion forums. Our analysis was based upon multiple offerings of a course on Big Data in Education. First we manually labeled forum posts into 12 categories (e.g. questions, answers, off-topic replies, etc.,) which are highly unbalanced and then developed a novel cascade model to automatically classify the comments by type. Second, we analyzed correlations between the number of student posts in different categories and their learning outcomes. We found that all types of forum contributions benefit learning outcomes. Finally, in order to support deeper analysis and to aid instructors in managing MOOC forums we applied Latent Dirichlet Allocation (LDA) an unsupervised modeling approach on the synchronous dialogues to identify which topics were being discussed throughout the course. Our main contributions in this work are: training an automatic classifier to organize unstructured forum text to support students' learning; designing a well-tested coding manual to categorize the student online posts; developing a cascade model that can be used to automatically classify the comments to organize the unstructured posts, even when the categories are imbalanced; and finally, we utilizing unsupervised topic modeling to provide deep insight into what is being discussed.

5.1 Introduction

As MOOCs develop rapidly, the number of participants and the amount of students' learning data can be massive. When people discuss a topic during collaboration, the amount of knowledge that they share is correlated with the amount that they learn [Reu15]. Indeed, one of the most important aspects of knowledge sharing is peer interaction [Reu01]. Prior researchers have shown that discussion in classrooms can facilitate student learning on course content and also student participation in online asynchronous discussions can enhance their learning [Coh94; Pal08].

However, due to the large scale and asynchronous character of MOOC discussion forums, they yield enormously unstructured content in the form of forum threads that reflect students' learning processes, their self- and co-construction of knowledge, their misconceptions, their opinions on the classes, their interests, and their learning outcomes. In a MOOC course, the discussion forum may be the only venue for students to collaborate, exchange ideas, and communicate with instructors. It may also be the only way for instructors to share supplemental material, interact with students, or provide help. If teachers and students can share supplemental knowledge or ideas with their peers, they can establish a virtual connection and co-construct new knowledge. Many researchers have investigated how the discussion forum connection is relevant to student learning outcomes [Bro15a; Jia14; Wis18]. However, they have not always drawn consistent conclusions about the correlation between observed student behaviors online and learning outcomes. One possible explanation for the phenomenon is they did not distinguish posts based upon its content. Thus, the overall goal in this work was to develop a tagging method and a foundational model to classify student posts and to help students and instructors to manage and benefit from MOOC forums. We will address the following three research questions.

- RQ1.2.1: Can we classify student posts by type?
- RQ1.2.2: What kinds of discussions are associated with learning outcomes?
- RQ1.2.3: What are the topics that students are discussing on the forums?

In order to answer *RQ1*, we first tagged the students' posts as described in Chapter 3 and then we developed a cascade classification model based upon support vector machines to automatically categorize different online discussion posts and comments. For *RQ2*, we highlight the relationship between forum discussions, i.e. the number of post types for each student, and the final student performance. Finally, for *RQ3*, we used Latent Dirichlet Allocation (LDA) topic modeling to identify which topics are associated with richer discussion, which can be brought to the attention of instructors and students.

The main contributions of this work are: first, the synthesis of a completely automatic process of organizing unstructured forum text to support students' learning; second, the development of a well-tested coding manual to categorize the student online posts; third training a cascade model to address unbalanced data classification problem; and fourth, the use of an unsupervised model to uncover 'hot' topics which students discussed.

5.2 Background

5.2.1 The ICAP Framework

In this work, we developed a coding manual to optimize the ICAP framework [Chi14] to better capture students' learning activities in MOOCs by reading all student discussion forum posts and studying the relation from their posts to their learning outcomes, i.e. assignments and quizzes. The framework proposes that all students' engagement behaviors can be categorized into one of four different modes with priority: interactive, constructive, active, and passive. *Interactive* discussion includes a clear exchange of information and co-construction of knowledge, e.g. asking and answering comprehension questions with a partner. *Constructive* knowledge includes generating information that is more enriched and coherent, and goes beyond present information, e.g. taking notes in one's own words. *Active* activities includes learners' focused attention engagement with instruction materials, e.g. summarizing by copy-and-delete. And finally, *passive* focuses on students receiving information or being oriented toward it without explicit engagement, e.g. reading entire text passages without doing anything else.

Previous studies of standard courses have sought to categorize students' engagement behaviors based upon their discourse. Marzouk et al. [Mar16] based their analysis on the ICAP framework [Chi14] which focused on the value of different dialogue contributions and their relation to students' learning processing. The ICAP framework has also been widely used in assessing student engagement in other venues. It has also been applied in MOOC forum activity analysis, since MOOCs provide a computer-based collaborative learning environment where large social connections may clearly reflect students' cognitive presence. Wang et al. applied a content analysis approach to analyze students' cognitive behaviours on the MOOC forums and explored how participation influenced their learning gains [Wan15a]. They extracted active, constructive, and interactive activities from ICAP framework as their classification categories. Then, they generated the top 10 features (words) for each category from their regression model. If a student used the top features in their post, it is likely that that post belongs to a certain category. For example, 'lecture' is the top feature for active activities, if this word shows up in a post, that post has a large chance to be an active activity. Consistent with the ICAP framework, they observed that students' *active* and *constructive* discussion forum activities were highly relevant to predicting their learning performance. They completed a follow-up study that further explored what kinds of discussion forum behaviours contributed to higher learning gains [Wan16a]. They found that students who exhibited more high-order thinking behaviours learned more and had deeper engagement with posts on the forum. They also applied a topic modeling method called Latent Dirichlet Allocation [Ble03] on the course materials. They concluded that socially-oriented topics are positively associated with richer discussion, which means more discussion in the forum, while professional-oriented topics are negatively associated with richer discussion.

For the purposes of this part of our research, we developed a novel annotation framework that extends prior ICAP models by adding additional categories. For instance, we divided asking and answering question activities(*Interactive*) by the content: course logistics, course material, and programming languages. Also, we added students' non-engagement behaviours into ICAP, such as politeness expression, and off topic. We describe the details of our coding schema in section 5.4.1.

5.2.2 Content Analysis

When compared with traditional classes, MOOCs provide information to a large number of participants due to free registration and a lack of formal pre-requisites. This, in turn, provides researchers with rich datasets about students' learning comprehension, level of engagement, and attitudes. Some prior researchers have applied NLP techniques to identify the relationship between student success and their activities on the discussion forums (e.g. [Cro16; Wen14]). Wen et al, for example, examined the content of forum posts in MOOCs that include students' attitude towards the course and whether they completed it. They found that student that posted messages with positive motivation words and personal pronouns had a lower probability to drop out of this course. They also reported a significant correlation between the number of daily drop out students with sentiment features [Wen14].

Lin et al. built a cascade SVM model which classified six categories of posts [Lin09]. They analyzed data from an online discussion forum in a Traditional Chinese Moodle used by a public high school on a course of earth science. In addition to individual words and terms extracted from the discussion posts, they also explored different kinds of contextual information retrieved from the discussion forum database. The dataset includes the sequence number or position of each post in the thread, the categories for the parent posts, the post length, and whether a post contains a question mark. They found that combining lexical and semantic features improved the performance of their classification model. Of the six categories in their results, the highest classification f-score was announcement class 0.717 and lowest was conflict class 0.132. They also revealed that standard classification algorithms for imbalanced datasets perform poorly [Lin09]. However, from their results, the performance of the smallest population classifiers still performed poorly.

Instead of analyzing students' social activities from forum discussions, other researchers have explored the impacts of their behavior in other parts of MOOCs. Nicholas et al. combined students' posts with their video clickstream and found that it is possible to build a feasible classification model which can be used on data from multiple courses [Ste17]. Tucker et al. explored the emotion extracted from forum content and its effect on students' learning outcomes [Tuc14]. They found that there was a slightly positive correlation between students' sentiment expression with their quiz scores but a strong negative correlation with their assignment scores. Crossley et al. examined the correlation between student forum content features, such as numbers of word types, average post length, word meaningfulness, etc., with whether they completed the MOOC [Cro15; Cro16].

Based upon Lin's cascade model, we optimized the approach of selecting the category for each cascade layer. We chose the order of the post categories for the cascade model based upon the corresponding classifiers' performance.

5.2.3 Topic Modeling

In order to support topic extraction we relied on LDA (Latent Dirichlet allocation) [Ble03]. This is a probabilistic topic model used to uncover latent topics in texts and has been widely used to summarize MOOC discussion posts based on the inferred topics. Thushari et al. [Ata16a] proposed a framework to auto generate and label MOOC discussion posts based on topic modeling. The authors extended the work to propose a topic visualization dashboard that can aid in understanding emergent discussion themes and identifying popular topics of discussion in MOOC forums [Ata16b]. They claimed that their topic modeling dashboard facilitated the instructors to locate the most influential discussion topic. This allowed the instructors to respond to those topics based the corresponding course material. Ezen-Can et al. [EC15] used clustering algorithms to group posts from MOOC discussion forums into clusters and then employed LDA to capture the topical theme of each cluster. They concluded that unsupervised clustering models can develop a framework for synchronous forum discussion. Topic modeling is also necessary to make the model more interpretable and cohesive. In contrast to their work, we extracted topics from each thread, which contains more than one post. Then we applied k-means clustering algorithm to organize thousands of topic words that we extracted from the hundreds of student dialogues.

Our reviews of the literature reveals several critical issues. First, with the widely used coding schema, course relevant versus non-course relevant, researchers are still not able to draw a consistent conclusion about how discussion forum activity contributes to learning outcomes. Thus, we developed a more complete coding schema to classify forum posts by adding more categories. For example, in the ICAP framework, *Interactive* indicates students asking questions for knowledge sharing. We divided asking question activities by the content: course logistics, course material, and programming languages. Second, different from Lin's cascade model [Lin09], we optimized the approach of selecting the category for each cascade layer. Finally, we applied topic modeling and k-means clustering to summarize the dialogue in the MOOC discussion forum.

5.3 Data

Table 5.2 presents the summary information about the forum threads and posts from both offerings of the course. Table 5.1 shows the number of posts and threads each user contributed. The table illustrates that the number of words in the posts vary widely, which indicates a diversity of post activities.

Table 5.1 Number of posts and threads per users

BDE2013	min	max	mean	median	std
# of posts	1	463	5.43	2	23.6
# of threads	1	277	3.25	1	12.3

BDE2015	min	max	mean	median	std
# of posts	0	637	3.97	1	28.4
# of threads	1	542	3.06	1	23.95

Table 5.2 Summary of posts and threads

BDE2013	min	max	mean	median	std
Length of post	0	1721	64.8	42.0	82.9
Length of thread	1	150	6.8	4.0	11.0

BDE2015	min	max	mean	median	std
Length of post	0	445	37.2	26	39.2
Length of thread	1	37	3.1	2	2.6

5.4 Methods

We developed a process to automatically organize the discussion forum. First, we annotated our datasets with our own coding schema. Then we built a cascade structure SVM with lexical and structural features. Third, we investigated the correlation between the number of different forum posts and students' learning outcomes. Finally, we applied LDA topic modeling methods on the thread level and clustered the topic words to make it possible for users to know what is being discussed by others without diving into the massive scale of discussion.

5.4.1 Coding Schema

Systematic direct observation provides one of the most useful strategies for accomplishing this goal by focusing efforts on socially significant and meaningful behavior change [Hin02]. Thus, in this work, we built our coding schema based upon the qualitative observation of the entire dataset.

We divided our coding schema into professional and non-professional topics. We defined professional topics as topics that are related to the course such as course content, software environment setup, and course logistic related. We defined non-professional topics as topics that are not related to the course, such as random chat, show politeness, etc. Our decision to segment non-professional posts was driven by prior literature which has examined the topic closely [Leh10]. Lehman et al. [Leh10] analyzed whether non-professional is a learning opportunity based on 50 hours of one-one expert tutoring sessions. They analyzed the occurrence of the off-topic word during the conversation. They figured out that tutors maybe use non-professional conversations to open the conversation and discuss more general studying strategies.

First, as for politeness posts, Peng et al. [Pen14] claimed that "the adoption of politeness strategies shortens the teacher-student social distance, makes the class interesting, and in turn facilitates English teaching and learning" . In addition, Bruce et al. [Ade16] present to determine how to improve the instructional effectiveness of an intelligent tutor system by focusing on the tutor's conversational style. In particular, they examined the impact of the conversation provided by the tutor whether or not using polite rather than direct wording of feedback and hints. They found that tutor's politeness conversation had a significant relevance with students' learning performance in ITS by ANOVA test. On the other hand, Adel et al. did a qualitative study to analyze politeness strategies in posts written by Iranian EFL learners in a class asynchronous blog. They found that the learners frequently use politeness strategies to post can help them build close relationships.

Second, as for social posts, Maher et al. [McL11] examined the role of student social chat in the online learning process. They found that social chat "allows for the development of related skills and knowledge and is an important precursor to more formal learning opportunities and as such, should be considered as an important part of the learning process".

Third, as for other random off-topic chats from students, Cade et al. [Cad10] categorized non-professional conversation into different categories based on the text content. And they found that different categories of non-professional conversation have different impacts on students' communication. Thus, we distinguished non-professional posts categories include: 'Politeness', 'Off topic', and 'Social', to dive deeply into how text content related to MOOC students' learning.

For professional posts, we first divided them into whether or not the post is a question. Cui et al. [Cui15] investigated how content-related questions in MOOC discussion forums related to students learning outcomes. They found that ask questions will help MOOC students get better performance.

Among professional topics, first, we developed on-topic(course content related) posts based on interactive activities. interactive activities defined by Chi et al. [Chi14] as exchanging information between students. For example, students asking and answering comprehension questions with a partner. Then, we developed software issues posts. The software issue also covered course material. It's necessary to course but not from the material as background knowledge. Finally, we defined course logistic issue posts as other questions related to complete the course but not related to course materials, such as when is the office hour.

- **On topic question (Q)**: Questions related to course topics or homework completion.
- *Example for Q*: Why would one use PivotTable on this problem when a mere two data sorts, one on each dataset, gets one to the answer?

- **On topic statement (O)**: Answers to questions or opinions on Q questions.
- *Example for O*: These are the coordinates of the centroids of each cluster. So cluster 0 is centered at [892.503, 513.662, 544.192, 461.702, 473.589, 866.709]. If you compare this to cluster 1, you note that there's a big difference between the centroids for attributes a and f, so those are the features that differentiate between clusters 0 and 1.
- **Question about technical issues (BQ)**: Technical questions about coding or software issues.
- *Example for BQ*: This site leads me to SourceForge which is the PC version. I could not find installation instructions for mac users.
- **Technical issues statement (B)**: Answers to questions or opinions on BQ questions.
- *Example for B*: This is indeed from the SimpleAPrime.java code in the computeAPrime.zip file from Dr Baker's site. I found the Wikipedia page afterwards - searching for Wilcoxon.
- **Course logistic questions(CQ)**: Questions about course logistics.
- *Example for CQ*: his problem seems to be worth 17%. I have gotten the other 10 problems correct, but my Assignment 2 score is 83%. (even after refreshing the browser page). Or is there more to Assignment 2 than these 11 problems?
- **Course logistic (C)**: Answers to questions or opinions on CQ questions.
- *Example for C*: This issue is still happening, this is quite worrisome...I got a perfect score on the first quiz and don't want to have to re-do it in the form of "Assignment 1B" or whatever before the Nov. 7 deadline. I'd prefer not to waste my time.
- **Announcement (A)**: Announcements to the class.
- *Example for A*: Hi everyone. All materials are released Thursdays at noon USA Eastern, so quizzes (assignments really) are also due at that time. (But two weeks later).
- **Additional materials (M)**: Provisions of supplemental materials or videos for students.
- *Example for M*: Rapid-I have video tutorials on their website.http://rapid-i.com/content/view/189/212/l
- **Social connection (S)**: Students' self-introduction to the class or attempts to set up discussion groups.
- *Example for S*: Hello all! I'm ANON, currently doing research in Natural Language Processing at ANON. I am looking forward to learning about EDM and applying it to education in under-developed regions. I will also be applying for graduate studies in educational technologies using NLP/Machine Learning. So, hope this course helps!

- **Politeness (P)**: Expressions of politeness or expressions of gratitude to classmates or instructors.
- *Example for P*: Thank you for adding this! Very helpful!
- **Off topic (T)**: Irrelevant posts or broad complaints about the instructor or the course.
- *Example for T*: usually, the class job is a routinary experience, the numbers are not a frecuently skill in our practices, and, if i dont make a measure, i cant achieve the goals.
- **Other language(X)**: Posts not in English.

Table 5.3 shows the distribution of items in each group for professional topics and non-professional topics. As the table illustrates, the final three categories (Additional Materials, Announcements, and Other Languages) were rare. We therefore elected not to develop classifiers for them. This distribution is illustrated in Figure 3.2 as well. We observed that our dataset has an imbalanced distribution. However, this could be due to the fact that we did not separate instructors' and TAs' posts from the original dataset.

Table 5.3 Summary of professional and nonprofessional forum topics for the two analyzed BDE classes.

Post Category	BDE2013	BDE2015
Professional Post Categories		
On topic statement (O)	1292	228
On topic questions (Q)	666	133
Software issues (B)	406	362
Course relevant (C)	205	100
Question of software issues (BQ)	201	161
Question of course relevant (CQ)	105	66
Non-Professional Post Categories		
Social connection (S)	335	381
Totally of topic (T)	494	208
Politeness(P)	383	399
Additional materials (M)	14	10
Announcement (A)	9	8
Other languages (X)	149	36

5.4.2 Lexical and Structural Features

To convert text into numerical features for classification, we extracted lexical features and structural features as follows.

5.4.2.1 Lexical Features

We segmented each post into words (unigrams) and bigrams where the value of each entry was its term frequency-inverse document frequency score (described below). Thus in lieu of a count for the number of times a given word appears in the post, we estimated the relative importance of it for the post.

5.4.2.1.1 Tf-idf

Term frequency-inverse document frequency (Tf-idf) is a weighting method that is used to estimate the relative importance of a term or bigram in a given document by dividing its relative frequency within the post by its relative frequency within the entire corpus. Tf-idf is a ranking metric that was developed for information retrieval applications and is frequently used in text summarization and classification [Ram03].

5.4.2.2 Structural Features

As in the discussion forum, participants posts are not independent. Instead, all posts in the same thread are considered relevant to each other because thread followers must read the previous post before they make comments. Thus, we explored semantic features to implement this characteristic of thread conversation. This includes: parent and child node reply relations and question marks.

5.4.2.2.1 Parent and child reply relations

In the discussion forum, students often.used posts to express their views and opinions. Participants express their opinions by replying to a previous post. So, the parent and child of a post are directly related to the post in question.We first ran one round of prediction based on the post's text alone. Then we generated the post's parent and child feature with the first round prediction results. We extracted binary features for whether the post has a parent post and whether it has a child. Also, binary features for what categories are the parent and child post, for example, if the parent belongs to 'S', we set the feature 'parent category S' as True, and the others as False.

5.4.2.2.2 Question marks

Previous studies found that in many cases, question marks are a very important punctuation feature when identifying question sentences [The10; Sta00; Pan02]. So we included the presence or absence of a question mark as a separate binary feature in our classification task.

5.4.3 Learning Models

Support vector machines are supervised machine learning algorithms that are used for data classification and regression analysis [Cor95]. A SVM model represents the examples as points in space, mapped so that the examples of the separate categories are divided by a hyperplane.

5.4.3.1 Cascade

Our initial experiments with a single SVM produced poor results. This was due in part to the presence of infrequent classes (e.g. class M), and the imbalanced distribution of the remaining data. In order to address these problems, we developed a cascade model. Figure 5.1 shows the workflow of our cascade algorithm. It is implemented as a binary structure, and the root is the input data of all the numerical post features. Then, to decide the order of classifiers in the whole model, such as which posts should be 'category 1', we selected the best performance classifier for each layer. To achieve that, first, we built the classifiers for each category with the training data (90% input data) with one vs the rest strategy. Then, we tested all of them on the testing data (10%). From these results, we extracted the best performing one as the selected classifier ('category i'), where i is the layer index of cascade model. Then, we filtered out the posts predicted as 'True' and passed the rest of the data into the following layers. The input of the next layer is, all input data except those post predicted 'True' from the previous layers. We then iterated the procedure until the last class was predicted. Since we removed the most confident records from the data during prediction, we believe that the data noise was reduced.

5.4.4 Post types & student performance

We measured the significant relevance from students' forum posts to their final performance by Kruskal-Walls(KW) test. Kw test is non-parametric rank-based similar to the common Analysis of Variance [Kru52]. For each student, we calculated the correlation between the number of posts for each category and their final grade. Also, as previous research found that more than 80% of MOOC participants receive a 0 grade, we executed the same analysis excluding 0 grade students.

5.4.5 Unsupervised Dialogue Model

To provide the insights into conversations among learners in a MOOC discussion forum, we first applied LDA topic modeling for the threads that contains more than two posts. Note that one post thread is not a conversation, so we excluded them. However, there were hundreds of threads in each MOOC forum, so even after we generated topics for each thread, it is impossible for users to go through all the topic words. Thus, we further investigated the topics with a clustering algorithm to gain the deeper insight into the dialogue content. Finally, we combined the topic modeling and the clustering algorithms to provide an understanding of MOOC forum posts.

5.4.5.1 Topic Modeling

The topic modeling algorithm captured the topic of the dialogue. Different from other sources of dialogue where the topics are more constrained, MOOC discussion topics range widely. These topics indicate the students' forum behaviours, such as what they are struggling with. Therefore, it is important to understand and provide automatic support to students and instructors. In this study, we used LDA (Latent Dirichlet Allocation) [Ble03] to capture the topics in each thread. After we obtained the topic for each thread, we trained a Word2Vec model [Mik13] on each dataset, separately, to convert the topic words to vectors for the following cluster task.

5.4.5.2 Clustering Algorithm

We applied K-means [Wag01] clustering algorithm with the vector of key words. K-means clustering algorithm is widely used for text summarization [Zha02; Wan08]. Given the number of clusters, it automatically clusters each data point into the closest cluster by calculating the distance to cluster center. In this study, we set the number to 8 because it was an 8-week course which covered different topics each week.

5.5 Results & Discussion

5.5.1 Research Question: Forum Post Classification

To address RQ, "*Can we classify student posts by type ?*", we developed a coding schema and built a cascade SVM model automatically classify student posts with a parent and child structural feature.

With the cascade model we applied linear SVM as the classifier model based upon parameter tuning with grid search strategy. To reduce the bias of the model, we applied 10-fold-cross validation. Then, we ran two rounds of the cascade model to generate the final prediction results. The first round of the cascade model generated the prediction results which was used to obtain the parent and child reply relations. We ran the second round to get the final results. In order to make the result robust, 10-fold-cross validation with stratified policy was applied, which balanced the proportion of different categories in each fold. We repeated the procedure five times and reported the average value as final results. In order to measure the performance of our cascade model and the structural features, we applied two baseline methods: randomly assigned labels at the same proportion for each category [Bal00], and one vs the rest SVM classifier. Since the Tf-idf matrix is very sparse, we experimented with the number of features using a grid search method, and we kept the top 5000 feature for our analysis.

Table 5.4 and table 5.5 show the F1 score for each label on BDE2013 and BDE2015 dataset separately. P + C means parent and child structural features. The bold number indicates the best F1 score for each category. B_random is our baseline model of randomly assigning labels to each post; B_SVM is the baseline model of single multi-class SVM; Tf-idf is the cascade model with Tf-idf features; Tf-idf+p+c is the cascade model with Tf-idf and parent and child features. When compared with the two baseline methods, our cascade model improved the overall Macro-f1 score on both dataset. With the parent and child structural feature, both model performance increased. This indicates that our parent and child features are able to capture the surrounding context information of posts to improve the model performance. In both tables, the bold texts indicate the best F1 score for each class, as we can see most of the best results are obtained from our Tf-idf model with parent and child features, especially the result on BED2013.

In addition, we applied chi-square and mutual information feature selection methods on the Tf-idf feature, however, they did not improve the performance but instead decreased it. This may be because they reduced the number of features too greatly.

Table 5.4 F1 score for different models in BDE2013

Class	B_random	B_SVM	Tf-idf	Tf-idf+P+C
O	0.32	0.70	0.66	0.72
S	0.08	**0.82**	0.76	0.79
Q	0.16	0.50	0.64	**0.66**
T	0.11	0.39	0.53	**0.57**
P	0.10	0.63	**0.69**	0.68
B	0.09	0.51	**0.60**	0.55
C	0.05	0.09	0.32	**0.33**
BQ	0.07	0.20	0.38	**0.39**
CQ	0.02	0.15	0.34	**0.40**
Macro-F1	0.11	0.44	0.55	**0.57**

5.5.2 Research Question: Student Performance Analysis

To address Research Question, "*What kind of discussions are associated with learning outcomes?*", we investigated the correlation between the number of each type of post and students' final grades, and we only considered students that have non-0 grades. Table 5.6 shows the KW test for each category. For BDE2013, all the P-values are statistic significant, which indicates there is a strong correlation between the number of each type of post to the students' final grade. As for BDE2015, the only exception is the statement of 'S' (social connection) which

Table 5.5 F1 scores for different models in BDE2015

Class	B_random	B_SVM	Tf-idf	Tf-idf+P+C
O	0.10	0.35	0.47	**0.52**
S	0.20	0.88	0.70	**0.74**
Q	0.07	0.27	0.48	**0.51**
T	0.09	0.33	0.42	**0.45**
P	0.19	**0.92**	0.89	0.89
B	0.20	0.43	0.54	**0.59**
C	0.02	0.14	0.21	**0.25**
BQ	0.06	0.43	0.46	**0.49**
CQ	0.05	0.20	0.34	**0.39**
Macro-F1	0.11	0.44	0.50	**0.54**

indicates non-0 grade students' social connection post did not help them learn. We also investigated the same correlation for all students including 0 grade students. We did not observe different results for BDE2013. However, for BDE2015 including 0 grade students, there were no significant correlation between 'B' (programming language) posts and final grades. This result is consistent with the observation that all 2015 students were likely to discuss this topic. Additionally, there is not a relationship between 'T' (off topic) category and the final grade but has significant relevance with the 'S' category. For all students in 2015, they preferred to chat about off-topics conversations in the discussion rather than greeting each other.

5.5.3 Research Question: Forum Dialogue Summarization

To address "*What are the topics that students are discussing on the forums?*", we summarized MOOC forum dialogues using topic modeling and K-means clustering algorithm. Table 5.7 provides the representative topic words for each cluster. Cluster 1 shows some of the course concepts. Cluster 2 shows students shared their social media account to each other. Cluster 3 expresses the programming languages or the software in this course. Cluster 4 shows the the question number and some concept, this may indicate students confused about which result they should return for these questions. Cluster 5 is characterized by students greeting from the class and introduce themselves. Cluster 6 shows students concern about the weekly quiz deadline. Cluster 7 shows how students received a certification. From the clustering results, we found that though we could not return the cluster of topic words relevant to each week corresponding course content, the cluster successfully gathered students meaningful discussion.

Table 5.8 shows the sample forum posts from each dialogue and their clustered top topic words course content concepts, such as 'pearson', 'overfit', which could attract the instructor's attention and encourage them to provide some additional clarification for these concepts. Clus-

Table 5.6 Relation between performance and post type for non0 grade

Post type	BDE2013		BDE2015	
	H-value	P-value	H-value	P-value
B	182.1	<0.001	11.3	<0.001
BQ	259.9	<0.001	44.9	<0.001
C	389.8	<0.001	121.2	<0.001
CQ	474.8	<0.001	101.9	<0.001
O	11.1	<0.001	62.4	<0.001
Q	14.9	<0.001	43.9	<0.001
S	280.9	<0.001	4.7	0.03
T	190.5	<0.001	48.8	<0.001
P	114.8	<0.001	86.9	<0.001

ter 2 was composed of social connection conversation words, such as summarized 'facebook' when students shared facebook group to other peers so that they can discuss offline. Cluster 3 extracted the programming language students used to solve the quizzes. For instance, we captured 'python' from the dialogue that people discussed which language is good to implement BKT model.

Table 5.7 Clustered topic words

Cluster number	Topic words
1	metric, pearson, dimension, test, overfit
2	facebook, youtube, linkedin, github, morgan, cui
3	sql, python, database, log, rapidminer, cost
4	q3, q4, q5, q7,q9 , return, result, accuracy
5	welcome, education, codeschool, bigdata
6	quiz, week, deadline, classmate, edm, publication
7	portuguese, shanghai, delhi, phoenix, address
8	certification, instructor, project, tutorial

Table 5.8 Sample forum posts from each dialogue and their clustered top topic words

Cluster Number	Topic Words	Dialogue
1	metric pearson dimension test overfit, ...	-"You can calculate;Pearson's correlation in RapidMiner by using the operator Correlation Matrix. (...) -I apply Linear Regression Model for RMSE Performance. I get 0.235 +/- 0.000 - (...) -"Thanks a lot. I just need a simple instruction on the how-to in Rapidminer.(...) -may be you can read this explanation at(...)
2	facebook youtube linkedin github morgan ...	-Hey guys , I just created a facebook group (...) -hii.. just started with data mining.. wanna learn more about it..??
3	sql python database log rapidminer cost, ...	-Is anyone else trying to implement BKT on their own? I have a Python implementation (...) -Hi David, I'm going to implement it in Java but I'd like to have a look at the Python (...) -Thanks for the positive feedback.;I do think more and more data will be available over time, and hopefully

5.5.4 Error Analysis

Although the cascade SVM with structural features benefited the forum post classification model, the performance is still not acceptable for further analysis. In order to understand the problems that lie in the model, we analyzed some of the miss classified sentences. As an example, from a long thread where students introduced themselves to each other, we observed a sub-sequence prediction label of {...S, S, S, S, BQ, BQ, S, S...}. In that sub-sequence, our model misclassified two continuous sentences into 'Programming language (BQ)'. This indicated that the parent and child feature did influence the classification, because the model recognized most of the continuous 'Social Connection' posts. However, the misclassification of the first 'BQ' post also affected the prediction of the following post. Though the parent and child structural feature showed impressive power on classifying structural sentences, in the future, we will consider more adjacent posts on the sequence of the thread posts label.

5.6 Conclusions

MOOCs provide opportunities to a large volume of people who have internet access. Understanding student forum activities helps instructors to provide real-time intervention to support and facilitate students' learning. In this study, we investigated how the textual information from

MOOC discussion forums affect student learning and we provided a procedure to automatically extract information from forum posts.

To better support instructors and students in MOOCs, we sought to classify MOOC posts to investigate what kind of knowledge co-construction processes are associated with students' forum activities, and what type of textual expressions highlight different kinds of discussion forum activities. We then built a coding scheme by optimizing the ICAP framework to categorize forum posts based on different types of shared knowledge. Based on that, we developed a cascade model to automatically classify student text expressions and extracted top phrases students used.

We created a coding schema, with in which, students have 11 different kinds of co-construction of knowledge. These include: course content questions/answers; programming language questions/answers, course logistic questions/answers, polite replies, introductions to each other, and random chats. According to our analysis, about 29.5% of 2013 and 48.7% of 2015 posts in a Big Data in Education MOOC, are considered off-topic, i.e. politeness, social connections, or off-topic. That means only around 60% of the posts in the BDE MOOC were directly relevant to the course's content. According to Wang et al.'s work [Wan15a], they found professional topics contain less discussion compared with daily topics based upon the regression model, because they observed that, in their discussion forum, professional discussions were grouped together. Their results were opposite of ours. The reason could be, in our dataset, students did not tend to have a long threads focused on professional questions. They were more likely to separate overall questions into small pieces, and started threads for each piece.

The follow up analysis for building an automatic classifier with machine learning algorithms shows that SVM was the best-performing model when compared to Naive Bayes, Decision Tree, Random Forest, and XGboost in the student text classification task. To deal with the imbalance of small population categories, we proposed a cascade model that achieved better performance. In addition, identifying parent and child reply relations improved the performance of all of the models, which indicated that in thread analysis, the relationship with neighbors is a significant feature for prediction.

We found that for BDE2013, there is a significant correlation between the number of each post type and student performance. However, for BDE2015, if 0 grade students are considered, all types except Software issue and off topic are significantly correlated to the performance. The results explain that students may not learn much knowledge from others by only discussing their code bugs. By excluding 0 grade students, we observed that social chat posts are not relevant to students' final performance.

Finally, we applied unsupervised approaches to generate topic words from the posts. The results of the topic modeling and the k-means clustering shows the potential of understanding synchronous students discussion without diving into the forum.

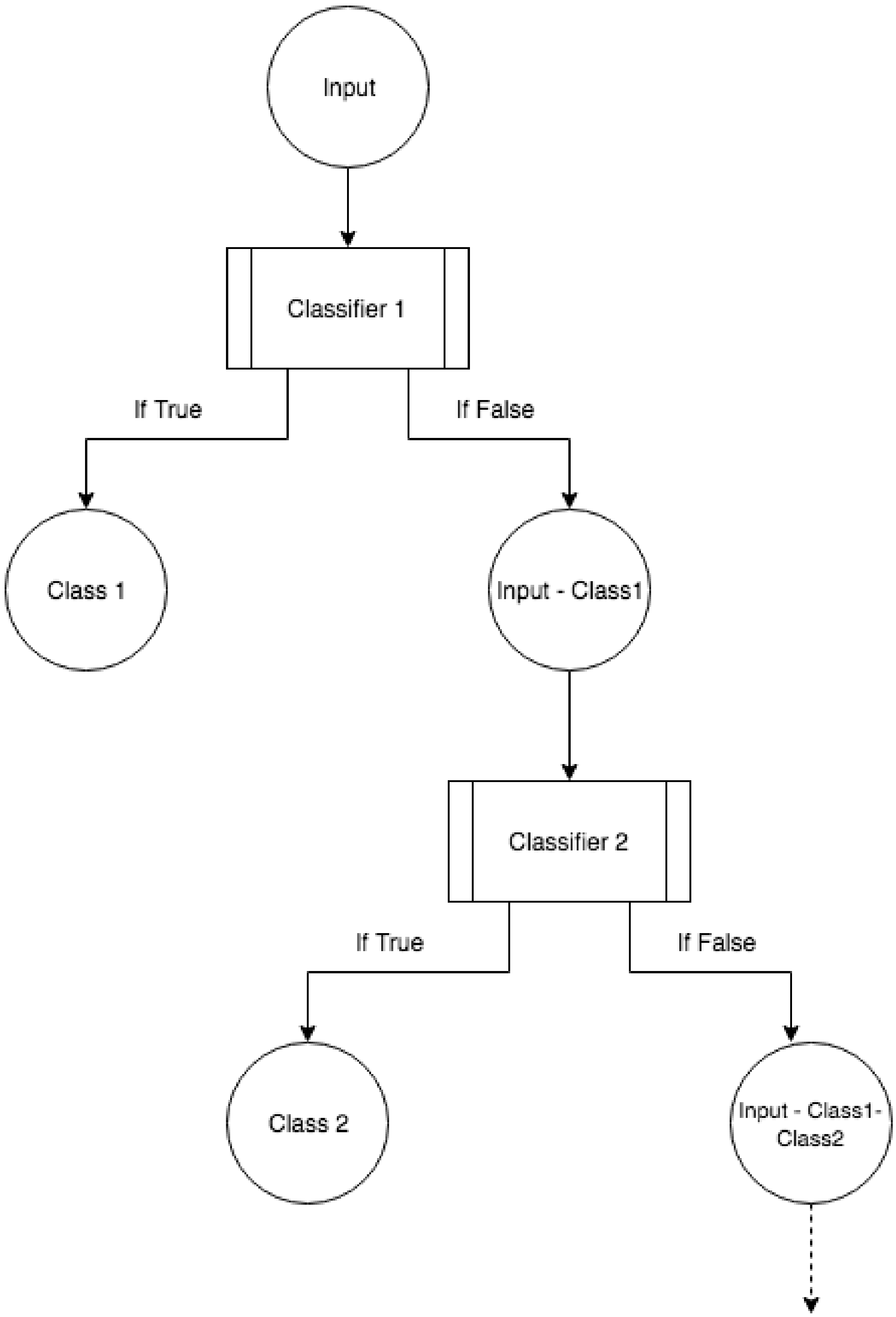

Figure 5.1 Cascade Model in each cross validation fold

Chapter 6

RQ1.3 - How do we analyze student online activities transaction combined their clickstream and forum activities?

MOOC Students' forum interactions provide researchers with insights into their online behaviors and learning habits. Prior studies have shown that these behaviors can predict MOOC students' dropout and their completion performance [Vu15; Wis16]. However, existing studies focus primarily on clickstream data. Most have not taken full advantage of the multiple types of data available including the semantic content from the discussion forum posts. In this work, we explore students' online platform behaviors based upon an educational data mining MOOC. We first examine how content from student posts relates to their final performance. Then we analyze their online transitions between problem-solving, watching lectures, and the different types of comments (e.g. questions, answers, off-topic replies, etc.) they made on the forum. The results indicate that students' post content relates to their performance and we are able to find significant differences between the online action sequences of high and low-performing cohorts. These findings will help MOOC instructors detect what behaviors contribute to student learning, what hampers their success, and provide guidance to students.

6.1 Introduction

MOOC Students' use of online tools, such as learning management systems, video lectures, and discussion forums, provides a vast amount of clickstream data on students' transitions within and between each tool. Such data can provide researchers with insights into students' engagement activities, their learning behaviors, and their learning habits. Prior studies have analyzed these data to investigate MOOC student performance, completion, and engagement

[Bal13; Boy15; Yu19]. Balakrishnan et al. [Bal13] found that clickstream features can be used to predict whether or not students will drop out in subsequent weeks. They achieved a high score (F1 = 0.89) at predicting students that stay in the class. However, the prediction score for students that dropped the course was very poor (F1 = 0.12).

The discussion forum is typically the only medium for students in MOOCs to collaborate, exchange knowledge, and communicate with instructors. As a result, students' forum discussions represent a rich record of their learning processes, their confusion, their opinions on the course, and their actions such as help seeking. Wise et al. examined the relationship between MOOC students' forum contributions and course performance [Wis18; Wis16]. They found that whether posts related to the learning of course materials is an important feature to distinguish higher and lower performing students.

Some studies that take advantage of both clickstream and forum data. In prior research we conducted survival analyses to predict student dropouts (ANON).We constructed social network representations based upon students' implicit social connections on the forums via posts and replies, extracted network features, and combined clickstream data to predict students' dropout. Crossley et al. [Cro16] previously combined clickstream data and NLP approaches to examine if students' online activities and the language that they use in the forums as related to successful course completion. However, they applied their analyses independently to the clickstream and forum activities and fed the results into their model rather than integrating activities across the social and study aspects of the course. This approach is typical of work in this area that has analyzed forum interactions or study behaviors, but not both.

Despite the substantial research on MOOC clickstream logs and discussion forum content, research in this area is limited and primarily focused on extracting features for performance prediction models. Our goal in this work is to analyze the relationship between MOOC students' online actions and their course completion. We address the following three research questions.

- RQ1.3.1: How can we categorize MOOC forum posts based upon their content?
- RQ1.3.2: Do MOOC students' discussion topics matter to their performance?
- RQ1.3.3: What are the most common transitions between MOOC students' online actions? And how do those transitions differ by student performance?

In order to answer these questions, we analyzed data from two offerings of a MOOC on Big Data in Education. To address *RQ1.3.1 & RQ1.3.2*, we first tagged comments and posts in the discussion forum into 12 different categories (questions, statements, off-topic, etc) based on their content. We then investigated the correlation between the number of students' posts in each type and their final performance. With respect to *RQ1.3.3*, in our previous work, we further studied the sequences of blended course student transitions between online tools and identified which learning activities made the most differences between the higher and lower performing cohorts [Git19]. We applied the same methods from our previous study to analyze students' action sequences [Xu19]. We selected completion as the evaluation metric in this study because it is one of the most common measurements for MOOC research [He15]. With the online activity information, we can understand asynchronous student online learning behaviours without diving into the forum and logs manually.

The answers to our research questions will allow us to better understand the patterns of MOOC students' online activities, to find the significant differences between higher and lower performing students, and help instructors identify likely harmful transitions to provide interventions to students. The long-term goal of this work is *to address the overwhelming and chaotic information in MOOC students' activities to support instructors' teaching and improve students' learning.*

6.2 Literature Review

6.2.1 Students' clickstream Analysis

MOOC clickstream data provide a vast amount of information including records of students' interactions with course materials (video lectures, assignment submission, etc) and discussion forums. Prior studies have investigated the potential of applying these data into student completion and engagement prediction [He15; Bal13; Boy15; Sha15; Tay14; Whi15].

Prior researchers have applied machine learning algorithms, such as logistic regression and hidden Markov models to predict whether MOOC students drop the course in the next few weeks. Balakrishnan et al. used Hidden Markov Models to understand student behaviours overtime by identifying students learning behaviour patterns from their interaction with platform, and predicting their drop out [Bal13]. They found that students who consistently watched more lecture videos(50%) were less likely to drop the course(0.1%) and who viewed at least a thread a week were very unlikely to drop(4%). Other researchers predicted student performance with their clickstream. Brinton et al. predicted student performance based upon whether a student will be correct on the first attempt [Bri15]. With the video watching clickstream data, they extracted summary quantities features (number of pauses, fraction paused) for the prediction model and outperformed their baseline methods. They also concluded that pausing videos will help students answer correctly on the first attempt quiz submission.

In addition to predicting students' performance, clickstream data also have been used to cluster students into pre-defined cohorts. Sharma et al. classified students into active students or viewers by their graded submissions [Sha15]. They defined students who actively participate in the course and complete assignments as '*active students*', and students who only watch the videos as '*viewers*'. Among active students, they also found that the distinction students get higher grades in their first attempt submission than normal and failed students. Thus, clickstream data has shown the potential for mining and modeling MOOC students' learning behaviours.

6.2.2 Forum Posts Content Analysis

When compared to traditional classes, MOOCs provide the opportunity to learn to a larger number of participants due to free registration and a lack of formal pre-requisites. This in turn provides researchers with rich datasets about students' learning comprehension, engagement, and attitudes. Prior studies analyzed post content with MOOC student performance. Wen et al. examined the content of forum posts in MOOCs that include students' attitudes towards the course and whether they completed it [Wen14]. They found that student posts with positive, motivational words and personal pronouns have a lower probability to drop out of the course. They also reported a significant correlation between the number of daily drop out students with sentiment features.

Instead of analyzing students' social activities from forum discussions, other researchers have explored the impacts of students' behavior combined clickstream data. Nicholas et al. combined students' posts with their video clickstream and found that it is possible to build a classification model which can be used on multiple courses [Ste17]. Crossley et al. examined the correlation between student forum content features, such as numbers of word types, average post length, word meaningfulness, etc, with whether they complete the MOOC [Cro15; Cro16]. Their results indicated that students who completed the course were those who interacted more with the platform more and participated more in the discussion forum. In addition, students who produced higher quality posts were more likely to complete the course. For instance, in terms of higher quality posts, students used more frequent words, more cohesion between paragraphs, greater use of connection words, etc.

6.2.3 Sequence Analysis

In addition to predicting student dropout, final grade, and completion with clickstream or forum posts, researchers have analyzed sequences of student actions from their logs. Several algorithms have been used to analyze student actions. One approach to analyze student sequence actions is considering the sequence as a sequence of strings and then identify the most common n-grams in it. Li et al. classified MOOC students into four achievement levels with n-gram features in the clickstream data and they claimed to outperform the state-of-the-art methods significantly with these sequence features [Li17]. Sinha et al. examined hidden structural configurations in MOOC students' learning activities. They combined video clickstream interactions and forum actions [Sin14a]. They generated n-grams from the combined actions to predict student attrition. They also found that these n-gram sequences improved their dropout prediction SVM model. Different from our study, they considered discussion forum activities as: post, comment, upvote, downvote, etc. Instead, we treated different types of post content as forum activities.

Another approach for analyzing sequences of student actions is by clustering them. Desmarais et al. collected the action logs of students in a college math learning environment [Des13]. They extracted student activity sessions from the logs. They first defined a 5 minutes cutoff time, if no event occurs in 5 minutes, the current session ended. They then clustered the sequences using the Levenshtein distance, an agglomerate hierarchical method which used to aggregate the most common similar sequences and identified three types of sessions. The first was when the students showed exploratory behavior and engaged in a mixture of browsing through exercises and notes. The second type was short sessions, comprising a variety of behavior such as browsing and attempting the exercises and quizzes. The third was exercise intensive sessions mostly consisting of exercise logs. Kizilcec et al. used a similar approach on student engagement in a MOOC [Kiz13]. For each assessment period, they labeled the students as either "behind", "on track", "auditing", or "out" based on their engagement with the course material. Then they applied K-means clustering on the sequences and were able to observe four clusters of students: completing, auditing, disengaging, and sampling. They found that students who say engaged through the course without taking assessments.

Different from previous studies, we focused on how different forum post content will impact students' learning. Most previous studies considered, however, we classified forum posts into different forum actions based upon the content such as related/non-related, question/answer, etc. And then we analyzed students' action transactions between different forum post actions and click actions.

6.3 Clickstream Data

clickstream data provided researchers with students' interactions on the platform during the course. In this study, we considered watching videos, chapters, and viewing pages as the same action: reviewing course materials(M). In addition, due to the differences of the platform, we were able to extract problem solving attempts(P) only in the 2015 dataset.

6.4 Methods

We began by annotating the student posts based upon our coding schema. We then investigated the correlation between the number of different forum posts and students' learning outcomes. Finally, we explored the correlation between students' sequential online activities and their final performance. We use the same coding schema in Chapter 5. Figure 6.1 shows the hierarchical structure of the coding schema.

Three experienced researchers annotated the two datasets using our coding schema. Two researchers annotated BDE2013 and two annotated BDE2015 with the lead author annotating both. We calculated inter-rater reliability using Cohen's kappa [Ber88] to evaluate the grader and manual reliability. Cohen's kappa is a measurement of agreement between two coders that accounts for chance agreement. Kappa represents the discrepancy between the observed probability of success and the probability of success under the assumption of an extremely bad case. Independence implies that the pair of raters agree about as often as two pairs of people

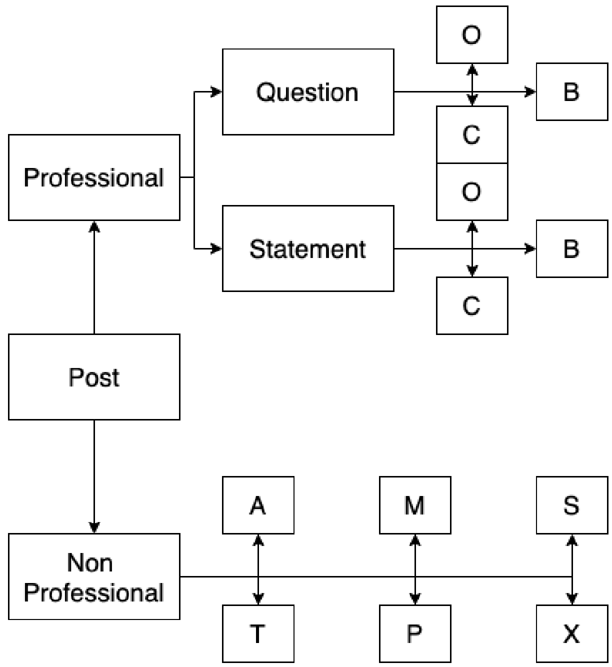

Figure 6.1 Coding schema

who make random decisions. The range for kappa value is from 0 to 1. In general, an agreement at 0.8 is considered to ensure high annotation quality, while sometimes a coefficient above 0.7 is good enough for text analysis [Ind10; Vie05]. To measure the inter-rater reliability, we randomly selected a 20% data sample from both datasets, and had three researchers independently label them. We obtained a 0.81 kappa for BDE2013, and 0.71 for BDE2015 which means that for both datasets, we maintain a good agreement when coding posts.

6.4.1 Sequence Generation

We combined the clickstream data and the tagged forum posts data. We then merged them into a single transaction data sorted by time. For each user, we then generated their study sessions based upon their online activities as in our prior study. Following Kovanovic's suggestion [Kov15], we explored our datasets to find the best cutoff time for generating students' learning sessions. If there were two continuous actions within less time than the selected cutoff, we considered them to belong to the same learning session. Otherwise, the previous session ends and the second action would be the beginning of a new session. In our previous study, we plotted the average number of actions per session, the number of total sessions based on different cutoff times. We applied the same method from our previous study, we chose session cutoff time at 7 minutes. However, MOOCs may have lecture videos longer than 7 minutes. Thus, for this study we investigated the time gap between clickstream actions and forum post actions. Figure 6.2 shows the results. We found that adding forum posts changed the nature of the individual sessions and forced us to reevaluate our cutoff. Our new cutoff value was now 20 minutes as the

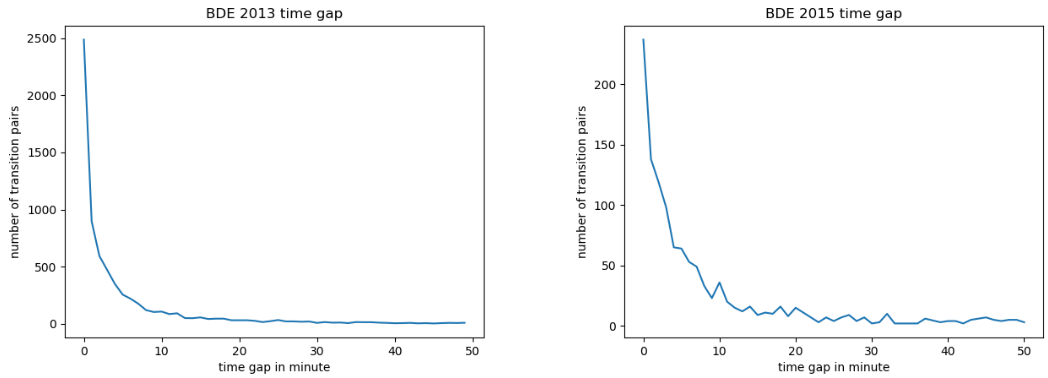

Figure 6.2 Time Gap between clickstream and forum post transitions

cutoff time. We selected this time gap considering students would post on the forum during or after lecture videos. Table 6.1 shows the number of students for each cohort and Table 6.2 shows the total number of learning sessions for each cohort. We found that students who completed the course contributed to most of the learning sessions.

Table 6.1 Demographic of Forum Users

	0 grade	non-0 grade	completion
BDE 2013	326	459	267
BDE 2015	362	156	65

Table 6.2 Demographic of Forum Users Sessions

	0 grade	non-0 grade	completion
BDE 2013	3326	24639	19754
BDE 2015	1349	2454	1639

We generated student action sequences from their learning sessions. We followed prior studies' rule to compact the action sequences by replacing continuous occurrences of the same actions with '+' notion [Mal11; Kin12]. We observed that most of the student sessions consisted of clickstream data. We considered all lecture video actions and page view actions as the same type of learning activity.

We used three different types of action sequences based upon student post type to investigate how students' posts were related to their learning process. First, we generated sessions only with the 9 post types(excluding posts in other languages, announcements, and additional materials) to examine how their forum discussions changed. Second, we generated sessions with students' clickstream and labeled post types grouped as "on topic" or "off topic" to examine whether the transitions between the different course platforms and different post topics differed among the performance cohorts. Finally, we generated sessions with students' clickstream and labeled post types as "question" or "answer" to examine whether student's help seeking or help providing behaviors influenced their performance.

6.4.2 Sequence Analysis

We applied n-gram analysis to identify the most common action sequences among students. N-grams represent a length of N words or phrases in a sequence in Natural Language Processing. They are considered important features, often used to convert text into numeric vectors. In this study, each student's sessions were transferred into sequence action words. We calculated the support for all sub-sequences of length $2 - 3$ to represent the transitions between two types of actions. We then collected these numbers for the distinction and non-distinction groups into two separate lists for each sequence. Prior researchers categorized MOOC students into 'active students' and 'viewers' based on their quiz submission action. In order to distinguish between different behaviours in the 'active students' group, we defined two kinds of distinction groups. First, students who completed the course by achieving a 70% or higher overall performance. Second, students who received credit for at least one assignment. We extracted the average support percentage for each sequence in each distinction group to find the most common patterns compared with non-distinction group.

Additionally, in order to compare the behavioural patterns that occur with different distributions among distinction and non-distinction student groups, we first calculated the average grade of each group and then applied Kruskal-Wallis (KW) test to see whether transitions are significantly different between the performance groups. Kruskal-Wallis test is a non-parametric method for testing whether samples originate from the same distribution [Kru52]. It is used for comparing two or more independent samples of equal or different sample sizes. A significant Kruskal-Wallis p-value between the student's behaviour transitions in two groups shows that the transitions are significantly different between the two groups.

6.5 Results

6.5.1 RQ1.3.1 & RQ1.3.2

Table 6.3 shows the proportion of posts in each category under our coding schema. As the table illustrates, the final two categories, Additional Materials (M) and Announcements (A), were comparatively rare and the teaching staff posted most of them. We were not able to interpret the

Other Languages (X) and therefore elected not to include them. This distribution is illustrated in Figure 3.2 as well. We observed that the distributions for the 2013 and 2015 offerings are different. This could be due to the fact that the EdX platform provides a private chatroom for students to discuss with each other. However, as mentioned previously, this type of data is not available to us in this study.

Table 6.3 Number of Posts in each category per class

Category	BDE2013	BDE2015
On topic questions (Q)	666	133
On topic statement (O)	1292	228
Question about technique issues (BQ)	201	161
Technique issues (B)	406	362
Course logistic questions (CQ)	105	66
Course logistic (C)	205	100
Social connection (S)	335	381
Politeness(P)	383	399
Off topic (T)	494	208
Other languages (X)	149	36
Announcement (A)	9	8
Additional materials (M)	14	10

Table 6.4 Correlation between post type and completion. '-': p-value > 0.1

	2013				2015			
	Non-0		All		Non-0		All	
	ρ	p	ρ	p	ρ	p	ρ	p
Tech issue	0.09	<0.05	0.20	<0.05	0.24	<0.05	0.37	<0.05
Tech question	0.07	<0.1	0.144	<0.05	0.13	<0.1	0.21	<0.05
Logistic issue	0.02	–	0.07	<0.05	0.12	–	0.18	<0.05
Logistic question	0.04	–	0.04	–	0.11	–	0.20	<0.05
Statement	0.22	<0.05	0.38	<0.05	0.21	<0.05	0.36	<0.05
Politeness	0.08	<0.05	0.23	<0.05	0.27	<0.05	0.28	<0.05
Question	0.14	<0.05	0.31	<0.05	0.23	<0.05	0.36	<0.05
Social	0.06	–	0.06	<0.1	0.02	–	0.16	<0.05
Off topic	0.03	–	0.01	–	0.21	<0.05	0.20	<0.05

Table 6.4 shows the Pearson correlation between the type of posts made by a student and their rate of course completion. We found that among non-0 grade students in both offerings, logistic issues or questions on the discussion forum did not help them achieve better performance. We defined logistic questions as questions concerning the deadlines, how to get certificates, etc.

We observed that among the MOOC students who tried to complete the course by doing at least one assignment, the ones that were more concerned about these types of issues did not receive a higher grade compared to the ones that were not concerned. In addition, for non-0 grade students, posts involving social connections did not benefit their learning. One explanation we have is that most of the social type of posts were posted at the beginning of the semester. However, as time goes by, some of the students in the MOOC lost their interest and drop the course. We observed that some of the results were not consistent among the two offerings. Similar to our previous analysis, one possible reason is that EdX platform provides a private chat to students.

6.5.2 RQ1.3.3

Since the platforms used in the two offerings differed, we could not collect the submission attempts for BDE 2013. Thus, we will present our results in each offering separately. We analyzed the sequence actions between different cohorts based upon students' grades for all discussion forum users. We investigated the relationship between the three pairs of student groups.

- All forum users who completed the course versus the ones who did not complete it.
- Non-0 grade forum users who completed the course versus the ones who did not complete it.
- All forum users who received at least one assignment grade versus users without any grades (0 grades).

Table 6.5 - table 6.10 shows the results of all pairs of cohorts in both course offerings. The numbers represent the average occurrences for each certain sequence of actions per user among the user group. And the KW p-value reflects the significant differences between the distinction and non-distinction groups. Note, these tables only show the transitions that are significantly different among the groups and the ones that are similar are not listed.

6.5.2.1 clickstream - On/Off-topic

In order to examine the importance of posting 'course related' and 'non content' posts, we combined post types into 'on topic' and 'off topic'. Politeness, Social connection, and Off topic types combined into 'off topic' actions and the rest were considered 'on topic' actions. The 'H value' in tables 6.5 - 6.10 represent the test statistic for the Kruskal-Wallis test [Kru52].

6.5.2.1.1 BDE 2013

Table 6.5 shows that no matter how we defined distinction groups of students, distinction students' clickstream transitions were much more frequent than non-distinction students' actions. This is not surprising due to the fact that higher performing students spend more time on the learning videos and page views.

Table 6.5 2013: Transitions between clickstream and on/off-topic action among all users(m: clickstream, t: on-topic, n: off-topic)

Distinct students by completion				
transitions	avg in distinction	avg in non_distinction	H value	Kw value
m+	69.1	16.3	427.21	0
m+t	6.0	2.6	63.09	0
m+n	2.5	1.89	10.30	0.001
mt+	10	1	4.00	0.04
mtm	2.6	1.3	4.80	0.03
Distinct students by 0 grade				
m+	51.1	10.8	411.36	0
m+t	4.7	2.8	52.26	0
m+n	2.26	1.92	24.42	0.000004
tm	4.4	2.5	45.30	0

Table 6.6 2013: Transitions between clickstream and on/off-topic action among Non-0 grade users

transitions	avg in distinction	avg in non_distinction	H value	Kw value
m+	69.1	25.2	229.74	0
m+t	6.0	2.6	23.27	0
mtm	2.5	1.89	5.67	0.001

Table 6.7 2015: Transitions between clickstream and on/off action among all users(m: clickstream, p: problem, t: on-topic, n: off-topic))

Distinct students by completion				
transitions	avg in distinction	avg in non_distinction	H value	Kw value
m+	9.5	2.8	152.63	0
pm	3.4	1.3	117.83	0
mpm	1.7	1.1	11.67	0.0006
np	4.2	2.4	11.70	0.0006
Distinct students by 0 grade				
m+	6.5	2.4	161.52	0
pm	2.6	1.2	84.84	0
mpm	1.5	1.1	4.72	0.03

With respect to the forum posting activities, distinction students had more transitions from lecture videos to forum posts, either to on-topic or off-topic posts. This indicates that, distinction students are more likely to have a discussion with other people when they learn new things or are confused. The results are also consistent with prior research finding that any kind of forum contribution will help MOOC students' learning. We also observed that the action sequence 'mtm: lecture - on-topic post - lecture' was significantly correlates to the students' completion rate. However, it was not a significant sequence action to distinct non-0 grade students. Instead, the sequence 'tm: on-topic post - lecture' was significantly different between the non-0 grade students and 0 grade ones. This might be due to the differences in students' motivations for registering in a MOOC. Some students were more likely to exchange their opinions in the forum before they dived into the lecture video.

Table 6.6 shows the results among the non-0 students. We found that any off-topic post transitions didn't have significant relevance to completion of the course. This result is interesting because, for students who are trying to complete the course, non-related posts in the forum do not help their learning.

6.5.2.1.2 BDE 2015

Table 6.8 2015: Transitions between clickstream and on/off action among Non-0 grade users

transitions	avg in distinction	avg in non_distinction	H value	Kw value
m+	9.5	4.1	81.08	0
pm	3.4	1.6	39.99	0
mpm	1.7	1.1	5.97	0.0006

Table 6.9 2013: Transitions between clickstream and QA action among all users(q: Question post, o: answer/statement post)

Distinct students by completion				
transitions	avg in distinction	avg in non_distinction	H value	Kw value
m+o	4.9	2.7	6.15	0.000001
m+q	2.9	1.5	17.56	0
qm+	2.5	1.5	12.60	0.000006
om+	4.2	2.4	3.93	0.00002
Distinct students by 0 grade				
m+o	3.9	3.8	4.45	0.0002
m+q	2.5	1.3	9.88	0
qm+	2.2	1.4	8.30	0.000001
om+	3.4	3.3	10.27	0.007

Tables 6.7 and 6.8 show the results of clickstream transitions with on/off topic post actions for BDE 2015. As we were able to collect quiz submission actions, we found that higher performing students had more transitions before and after they made a submission. We only observed that students will post off-topic content in the forum before they solve a quiz. However, on-topic post action sequences didn't have significant relevance with student performance. This might be because of Edx's private chat rooms where a separate part of student discussions might take place. As a result, our data on the student communications on the forum is not all the student communications in this course.

6.5.2.2 clickstream - Question/Answer

In order to examine the importance of seeking help and providing help, we combined post types into 'question', 'answer' and 'off-topic' as the same definition in previous section.

6.5.2.2.1 BDE 2013

Table 6.10 2013: Transitions between clickstream and QA action among Non-0 grade users

transitions	avg in distinction	avg in non_distinction	H value	Kw value
m+o	4.9	2.0	7.49	0.0005
m+q	2.9	1.7	7.74	0.0007
qm+	2.5	1.6	4.90	0.002
om+	4.2	1.8	5.82	0.0007

Table 6.9 and Table 6.10 show the student transitions between the clickstream actions

and their question/answer post actions. We found that distinction students always had more transitions between post and lecture videos. In addition, distinction students had more actions posting answers/statements than questions in the forum before or after they watched videos.

6.5.2.2.2 BDE 2015

We didn't observe significantly different transitions from the clickstream actions to question/answer post actions for BDE 2015. Instead, we found that non-distinction students had many more transitions from off-topic to post answer on the forum. This result is consistent with prior research finding that a set of MOOC users register as 'problem solvers', they were very active in the forum but they didn't try to complete the course[Sha15].

6.5.2.3 Post type

We were not able to identify transitions between different post types that were significantly different across performance groups. We analyzed the students' forum post type sequences in 20 minute sections, and the only significant results among both datasets was that for non-0 grade students, students who completed the course had more off-topic to off-topic transitions. This result is consistent with Wise [Wis18] finding that any forum contribution will help students in completing the course. The potential reason for not finding patterns in students' transitions from one post type to the other can be the 20 minute cutoff time for generating students' learning sessions. The students often took more than 20 minutes to exchange their opinions on the forum and come back for a new post. The median time gap between two on-topic comments for BDE 2013 students is 186.5 minutes and 76 minutes for BDE 2015 students.

6.6 Conclusion & Discussion

In this study, we combined logs from students' study behaviors using online platforms and discussion forums in two offerings of the "Big Data in Education" MOOC. First, we built a coding schema to categorize forum posts based on different types of shared knowledge in which students expressed 12 different kinds of co-construction of knowledge. These include course content questions/issues, technique questions/issues, course logistic questions/issues, polite replies, social connections, off topic, and other. We then analyzed the correlation between the students' posts by type and their course completion rate. Finally, we extracted student learning sessions from their activities and investigated the sequences of actions in these sessions to find general online learning patterns and patterns which distinguished higher and lower performing students.

Our results show that among non-0 grade students, logistic question/issues and social connection posts did not help students perform better in the course. In addition, for all forum participants, we found that distinction students always made more clickstream actions and they were more likely to transition between video lectures and forum posts. Interestingly, among non-0 grade students, posting 'non related' posts didn't help them complete the MOOC. Furthermore, among BDE 2015 data, we observed non-distinction students have much more transitions from off-topic post to post answers in the forum. However, within a 20 minute learning session, we did not observe transitions between post types which had significant relevance to students' performance. This result indicates that students may not post several different topics in a 20 minutes study session.

Considering the findings of this study, researchers and teaching staff can understand more about MOOC students' behavioural patterns. Once teaching staff identify good forum behaviours, they are able to encourage students to follow these patterns in order to succeed in this course. Furthermore, at current circumstance and due to the impact of Covid-19, more and more classes turn to online mode. Learning analytic methods and findings on MOOC students can also be extended and generalized to other online courses. One of the biggest challenges of the current studies to improve the impact of learning analytics is making findings more generalized with more data. There are large groups of researchers working on analyzing students forum behaviours. However, some of the conclusions from different studies on different datasets are even conflicting. So in future, by more collaboration among researchers in different groups, more data can be included in studies, which can help to make more generalized conclusions.

Chapter 7

RQ2 - How can instructors monitor students' learning status based on their use of the forum?

To answer research question 2, I explored using topic modeling to support instructors in monitoring MOOC discussion forums and developed a dashboard for instructional staff to monitor students forum activities. Summarizing the massive discussion threads into topics can provide an overview of discussion and support instructors in monitoring students' learning status without diving into the massive forum.

7.1 Introduction

MOOC classrooms are typically structured around a common set of educational tools including a basic learning management system, video lectures, quizzes, and a discussion forum [Shi15]. Students' use of these tools provides a vast amount of clickstream data on student transitions within and between each tool. Such data can provide researchers with insights into students' engagement activities, their learning behaviors, and their learning habits. Previous studies have shown that this information can help investigate MOOC student performance, completion, and engagement [He15; Bal13; Boy15; Hal14a]. On the other hand, the discussion forum is typically the only medium for students in MOOCs to collaborate, exchange knowledge, and communicate with instructors. As a result, students' forum discussions represent a rich record of their learning processes, their confusions, their opinions on the course, and their actions such as help seeking.

Searching and navigating massive forums is a challenging task for teaching staff [Ram14a]. To address this, most forums provide searching, discussion thread labeling, and some use the number of students' 'up-votes' on popular posts to represent their importance [Coe14]. These strategies can be helpful if an instructor is searching for a particular type of question/comment, but in many cases instructors read forum posts to gain a general sense of what their students are discussing. This is an important goal for instructors as they try to address the particular needs and interests of their students. Instructors might want to know if students are struggling with the content, how they are discussing ideas, or provide support to encourage students to work collaboratively. This serves the main intent of forums as a place to discuss ideas, course content, ask questions, answer problems, and build networks.

Currently if an instructor wants to know about forum discussions they need to navigate large quantities of posts, but this is overly burdensome. Topic modeling offers a novel way to reduce this burden by grouping similar posts together by topic and providing a topic description of their content. Grouping posts by topics for instructors can potentially provide holistic insights into the types of discussions occurring. Prior work using topic models on MOOC forum data has mostly taken a researcher perspective to understand discussions [Che16]. In addition, prior research analyzed forum discussions based on past post records [Che16; Pen20].

I did the following work to answer the research question "How can instructors monitor students' learning status based on their use of the forum? And How can we use our knowledge of structured forum to help instructor make initial interventions and improve students learning experience and students receive question responds more efficiently?". I first explored using topic modeling to summarize weekly topics for MOOC instructors. Then, I developed a dashboard that includes 'hot topics' and corresponding thread information for the instructors.

7.2 Literature Review

Visual analytics within the educational context often facilitate educators in understanding large amounts of learners' data to make inferences. For example, Akintunde et al. [Aki21a] developed a dashboard to analyze and visualize students' forum activities from a public discussion forum for a blended course. Within the MOOC context, detecting topics from an online forum is an efficient method to monitor the state of the forum to identify emerging trends. Topic modeling has been widely used to identify issues and people's opinions from text documents. Analyses were conducted to identify patterns within different professional domains such as health, education, finance, etc, from different kinds of data sources, communities, and locations. On the other hand, LDA (Latent Dirichlet allocation) [Ble03], a probabilistic topic model used to uncover latent topics in texts, has been widely used to extract topics from different data sources.

Prior studies build applications to support instructors teaching with topic modeling. ForumDash, a preliminary work by Speck et al. [Spe14], focuses on visualising which students are contributing, struggling, or distracted in order to facilitate instructors in targeting their efforts effectively, saving time managing online courses. Using three visualisation tools, ForumDash attempts to provide insights for teachers on which students contribute the most to discussions (i.e. Thought-leaders), identify topic clusters to determine popular topics, and through a 'contribution score visualisation' where students can monitor how much they are contributing to discussion forums compared to their peers. KISSME (The Knowledge, Interaction and Semantic

Student Model Explorer) [Tep11] is a visualisation framework that analyzes online discourse with the aim of understanding the nature of interactions among learners including contributions and relationships using LSA and social network analysis participants.

Prior researchers have also developed visualization dashboards and recommendation materials for forum questions. Ruiperez et. al. [RV15] developed a visualization tool that provides an extension of the learning analytics support for the Khan Academy platform. This tool includes new visualizations for the individual and entire class. It provides students with their learning progress for exercises and videos, i.e. how many quizzes and exercises students complete. Vigentini et. al. [Vig17] built a dashboard that provided quick insights for MOOCs. This dashboard provides the general class information like the total number of students who registered for and engaged in the course. It also shows the progress of each assignment for the entire class. In addition, Agrawal et. al. [Agr15] presented a tool named YouEDU in order to automatically detect and address confusions in forum posts. Leveraging the Stanford MOOCPosts corpus, they first trained a set of classifiers to classify posts into different dimensions. They then applied information extraction methods to map the confusion posts to minute-resolution clips from course videos.

Thus, our motivation for developing this study occurs due to a lack of research to provide weekly hot topics to analyze the overwhelming abundance of MOOC discussion content. As the discussion forum is the only official venue for students to communicate, weekly topics provide information that reflects students' weekly learning statuses, such as their questions and confusions. Different from prior research, my work focused on dynamic forum posts (weekly) to provide intermediate insight for instructors. I extracted topics by first extracting hot topics based on different weekly time windows, including topics based on the current time the instructors logged into the dashboard. Likewise, my recommendation module contains more auto-generated resources, including course materials and similar thread information.

7.3 Methods

7.3.1 Topic Modeling

We began by annotating the hot topics from the evaluation set. Annotations were made by 2 researchers. This process generated 2 human annotated topic sets for each week. In order to generate topics from the MOOC discussion content, we used Latent Dirichlet Allocation(LDA), which is a widely used topic-modeling method for exploratory analysis of large collections of textual data [Ata16a; Ata16b; Tep11]. LDA is a generative statistical model that considers each document as a mixture of underlying topics. A topic is a concept or theme that contains words that frequently occur together in the dataset and are used by the LDA model for learning the topics. LDA tries to identify these topics iteratively based on the co-occurrence of words in documents and represents each document as a composition of different topics with associated weights. In addition, we experimented with LSA [Ste07] and Non-Negative Matrix Factorization

[Lee99]; however, the results were worse when compared to LDA. Thus, in this study, we chose LDA as the algorithm to extract topics from MOOC discussion forum.

Since LDA is an unsupervised method (i.e., the data are not hand labeled), it is difficult to judge the quality of topics identified by the model. Measures such as perplexity or probability of held-out documents have been proposed for evaluating the quality of topic models, but they have not been found to correlate well with human judgment of topic quality. We evaluated topic modeling results by comparing the topics that were machine generated and human annotated.

7.3.2 Topic Model Evaluation

Evaluating the results of automatically generated topics against human judgment is a challenge [Mat19]. Statistical topic models have difficulty evaluating their outputs against concepts of human judgments [Cus15]. Statistical models often focus on statistical measures of perplexity or held-out likelihood. These features evaluate the predictiveness of the model given new data but do not measure the human interpretability of topics. Thus, the other option for measuring topic results quality is to utilize human experts to perform a human evaluation study [Cus15].

Based on previous work [Hag18], with rigorous training and evaluation, 87% of LDA generated topics made sense to human judges. Palese et al. [Pal20] showed how to evaluate topic modeling results with human annotation. This method measures the degree of human interpretability of the extracted topics. These topics are evaluated against a gold standard set of manually labeled documents. The gold standard set requires human assessment and identification of the topics in the sample documents.

A gold standard dataset for topic modeling in the educational domain requires educational domain knowledge from the coders so they clearly understand the definition of each topic. In addition, because each topic relies on contextual information from adjacent sentences, human topic identification is best performed at the document level. Thus, coders may see different topics from the same document.

The steps to create a gold standard dataset are:

- Corpus and topic selection: We have 16 weeks of forum posts from two course offerings. From this dataset, I randomly select 30% of the documents, or 5 weeks worth of forum posts.

- Introduction to the coding procedure: I manually coded the randomly selected documents with another educational domain researcher. In this procedure, I introduced the definition of the 'hot' topic to the other coder and answered questions about the definition and examples.

- Independent label: Each coder labeled their selected documents with the definition of the hot topic from each document.

- Reliability analysis: I used Cohen's kappa value to evaluate the reliability of the topics.

This approach is a direct response to measure how closely the topic modeling results approach human topic assignment. In summary, the steps of evaluation become:

- Create a gold standard dataset by human labeling of a random subset
- Run topic models with different hyper-parameters
- Compute models' accuracy by comparing with the gold standard

Based on these guidelines, I produced two gold standard datasets: partial agreement and inclusive set. The partial agreement set contains topics and documents with the same set of topics by different coders. The inclusive set contains all the documents and any topics identified by any of the coders.

7.3.3 Post Summarization

Once we extracted the topic keywords, we found the original threads and applied a text summarization algorithm to shorten posts for instructors to review. Text summarization is a technique to convert a long piece of content into a shorter one without removing the actual context. With the help of this technique, a summary of any text material could be generated. It only removes text data which does not change the overall meaning of the content. In this work, we summarized text using Term Frequency - Inverse Document Frequency (tf-idf), which is a widely used methodology in the text mining domain [Chr16; Pra16; Kha19]. We first preprocessed the text by removing English stopwords and removing symbols. Then, we tokenized the sentence and calculated the term frequency for each token. Finally, we calculated sentence scores based on the weighted term frequency and displayed several sentences that contained higher scores.

7.3.4 Extract New Questions and Recommend Materials

We extracted new questions by looking at whether the question had teaching staff replies. If no instructors replied the the question, we extracted this question to the dashboard to inform the instructor about the situation. With the extracted questions, we applied widely used text similarity algorithms [Che11; Alb14; Wan12] to find similar previous threads and corresponding course materials including textbook chapters and video lecture titles.

7.4 Results

We report the results as the list of keywords for each topics for each week. Among all of the figures, the color green represent overlapping of two different resources.

7.4.1 Difference between Coder1 and Coder 2

Figure 7.1 shows the comparison between two human coders. In this process, two human coders independently coded the hot topics from each randomly selected five weekly forum posts. The green background indicates the overlapping topics. From the figure we found that although most of the human selected topics overlapped, there were a few that were different between two human coders. This was caused by a difference of understanding of the forum by the coders. That is the coders identified the same 'hot' threads but with different topics.

Week 1	week 2	week 3	week4	week5
Zhikai				
Coursera enviorment	Course syllbus	social and networking	code sharing	CTAT environment tech issue
Coure success suggestion	social and networking	resource recommandation	CTAT assignment	welcome to class
Social and networking	QUiz 6	factor analysis	tools and enviorment troubleshoot	course suggestion
resource recommandation	reading csdv	Q1	resource recommandation	Self introduction
Quiz q1	resource recommandation	Q2	probability calculation	Question 1
Quiz q2	Confidence,Cosine and Lift	Q3		apperciation
Quiz q3	GSP dataset processing	Q5		problem clarification
Quiz q5	k-means	network data mining methods		
Quiz q6	tools and enviroment	graph mining tools		
R code troubleshoot	BIC formula	Quiz 8 tech issue		
RadpidMiner	EDM	clarification of lump		
Java configuration				
Dataset seeking				
Probability caclulation				
Data processing troubleshoot				
sharing code				
coding environment				
Yiqiao				
Complaint on the video	Probobility	Social introduction	quiz problem	CTAT assignment
Social introduction	Quiz 6	Quiz 8	CTAT assignment	social introduction
statistic	csv	factor analysis	probability calculation	assignment issue
Data mining	confidence	social network analysis	politeness thank	browser issue
Probobility	BIC	build network with thread	number of rules	welcome to the course
Quiz	dataset processing	software - graph tool		politeness thanks
RapidMiner	ethical question	other papers		
JAVA	cluster			
Educational papers				
BKT				

Figure 7.1 Coder 1 vs Coder 2

7.4.2 Difference between Human Coders and Topic Modeling

Figures 7.2 and 7.3 show the comparison between two human coders of topic modeling. Figure 7.3 shows the difference between Coder1 and Topic Modeling and Figure 7.2 shows the difference between Coder2 and Topic Modeling.

From the results, we could observed that most of the automatically generated topics were found in human extracted topics. However, machine learning generated topics often missed some of the human extracted topics. The might be due to generated keywords LDA produced for each topic. However, those key words may have also been too vague for humans to understand. For example, it is hard to understand what forum post containing 'Thank Luc question answer' means. Both human coders considered this sentence as a social introduction, as the student tried to show their politeness to another person. However the LDA algorithm only extracted 'question' from this sentence which is different from what student expressed. Based on these results, we decided to provide instructors with text summaries of the original header posts.

Week 1	week 2	week 3	week4	week5
Yiqiao				
Complaint on the video	Probobility	other papers	quiz problem	CTAT assignment
Social introduction	Quiz 6	Quiz 6	CTAT assignment	social introduction
statistic	csv	factor analysis	probability calculation	assignment issue
Data mining	confidence	social network analysis	politeness thank	browser issue
Probobility	Question 7 and 9	build network with thread	number of rules	welcome to the course
Quiz	dataset processing	software - graph tool		politeness thanks
RapidMiner	ethical question			
JAVA	cluster			
Educational papers				
BKT				
Machine				
course video coursare	BIC csv	Gephi cluster analysis	CTAT assignment	assignment
java question computeKTparamsAll	consine number concentrate value	link post thread graph	question answer calculation	welcome to the class
quiz	student observation raw	quiz	question hint thank try	browser cookie
BKT	curve learning time		number of rules	CTAT issue
probability model	thank Luc question answer		welcome to class	class glad think late
file column value csv	correlation support			
answer question thank datum	thread thread_id big data			
observation file column				
Feature cat meta valuing download				

Figure 7.2 Coder 2 vs Topic Modeling

Week 1	week 2	week 3	week4	week5
Zhikai				
Coursera enviorment	Course syllbus	social and networking	code sharing	CTAT environment tech issue
Coure success suggestion	social and networking	resource recommendation	CTAT assignment	welcome to class
Social and networking	QUiz 6	factor analysis	tools and enviorment troubleshoot	course suggestion
resource recommandation	reading csdv	Q1	resource recommandation	Self introduction
Quiz q1	resource recommandation	Q2	probability calculation	Question 1
Quiz q2	Confidence,Cosine and Lift	Q3		appericiation
Quiz q3	GSP dataset processing	Q5		problem clarification
Quiz q5	k-means	network data mining methods		
Quiz q6	tools and enviroment	graph mining tools		
R code troubleshoot	BIC formula	Quiz 8 tech issue		
RadpidMiner	EDM	clarification of lump		
Java configuration				
Dataset seeking				
Probability caclulation				
Data processing troubleshoot				
sharing code				
coding environment				
Machine				
course video coursare	BIC csv	Gephi cluster analysis	CTAT assignment	assignment
java question computeKTparamsAll	consine number concentrate value	link post thread graph	question answer calculation	welcome to the class
quiz	student observation raw	quiz	question hint thank try	browser cookie
BKT	curve learning time		number of rules	CTAT issue
probability model	thank Luc question answer		welcome to class	class glad think late
file column value csv	correlation support			
answer question thank datum	thread thread_id big data			
observation file column				
Feature cat meta valuing download				

Figure 7.3 Coder 1 vs Topic Modeling

7.4.3 Dashboard

Figure 7.4 shows an example of the dashboard. We provide two inputs for the instructors evaluating the dashboard. The first is the category of different post topics. Based on our previous work, we divided it into three different categories: Logistic, Technique, and Course Content. The definition of these categories can be found in Chapter 5.2.1. The second input is time window, where instructors could select between the current week's forum data and the last week's forum data.

The body of the dashboard is divided into two major parts. The first part includes three

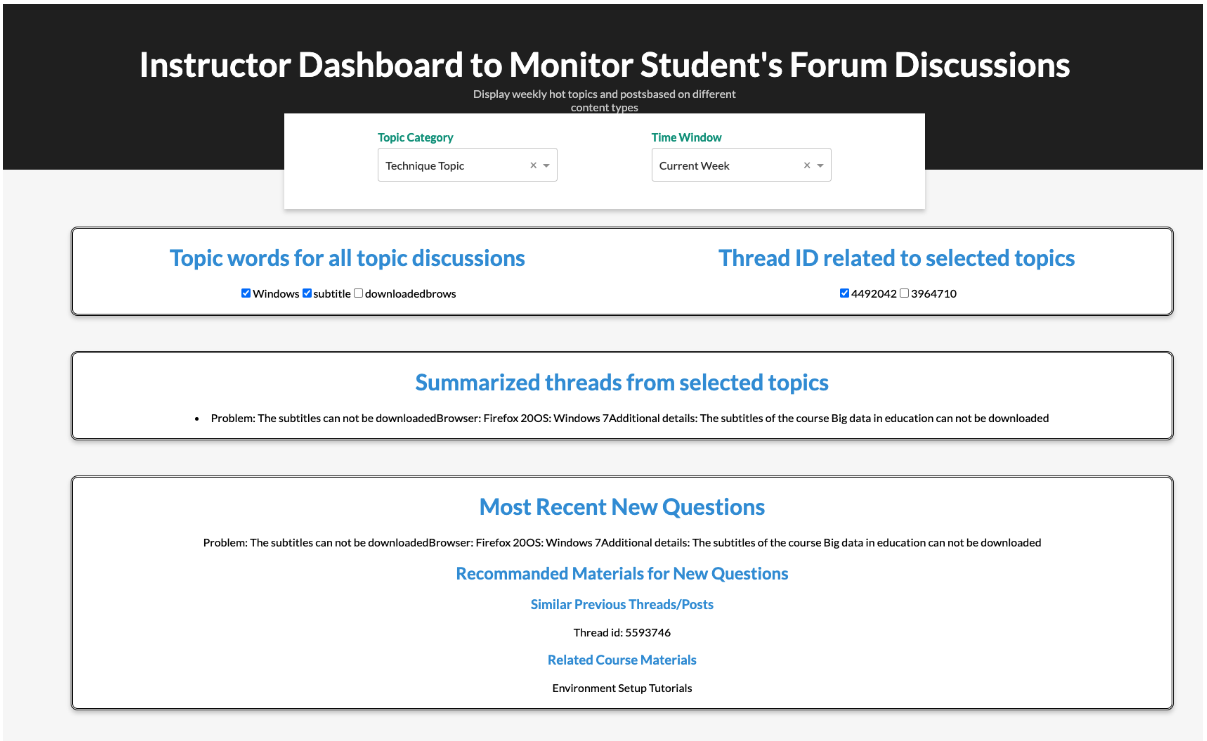

Figure 7.4 UI of the dashboard

sections: topic words, thread id, and summarized thread. The topic word section provides a list of checkboxes which contain the topic words extracted from the LDA topic modeling algorithm. If a user clicks on the word checkbox, the thread id section will display a list of thread ids which relate to the selected topic words. If the instructors wish to see summarizations of threads related to the thread id, they can click the thread id checkbox. The summarized thread information will display in the section below. The summarized thread information comes from the summarized first post of the thread. If the first post of the selected thread is longer than 10 sentences, text summarization is applied to summarize it to 10 sentences.

The second part of the dashboard displays the unanswered questions related to selected topic and recommended materials including similar threads and related lecture video titles. The most recent new questions are extracted based on the input time window and category within question posts. With this information, the instructor can learn what are students' common questions, allowing them to then prepare interventions, such as a Q&A section to answer the common questions and to save their time answering similar questions. If the logistic category is selected, the dashboard will display a message to remind users to check the course syllabus or direct email to the instructor staff. Otherwise, the dashboard will show the related thread id and related course materials based on the text similarity algorithm. We evaluated our dashboard by interviewing university instructors. The results are reported in the Chapter 8.

Thus, we answered Research Question 2 by first, evaluating the Latent Dirichlet Allocation topic modeling algorithm, and second, developing a dashboard with extracted hot topics and recommended course materials related to the selected hot topics. We found that LDA is the current best algorithm to extract topics from text documents and monitor student forum discussions. In addition, based on the results of the instructors' survey reported in Chapter 8, our content-based dashboard could improve instructors' teaching experience.

Chapter 8

RQ3 - What are instructors' goals in monitoring student interactions? How do they evaluate adaptive scaffolds? What features do they want in an adaptive scaffold for content monitoring?

In this portion of the work, we investigate how automated structures and content-based scaffolds can support classroom orchestration and learning by surveying MOOC and university instructors. We sent out survey requests and eventually received responses from 4 MOOC instructors and 10 university instructors. Further, we interviewed five of the university instructors who were willing to speak with us in detail about their teaching experience, teaching philosophies, and their opinions on the dashboard designed in Chapter 7.

8.1 Introduction

The majority of prior research on MOOCs has been focused on investigating students learning motivations and behaviour patterns. Little attention has been made to MOOC instructors who organize the course and make a MOOC happen. Bonk et al. [Bon18] proposed a study of MOOC research publications between 2014 and 2016 and showed that instructor-focused research is the least rank (3.4%) after student-, design-, context-, and impact-focused research. Some researchers have investigated how course designers construct their MOOCs [Ber08]. For example, Mak et al. [Mak10] explored how the MOOC instructors intended to solve the high dropout rate problem by summarizing ongoing discussions in the forum and sending a newsletter to students. Thus, it is important to understand MOOC instructors' workflow and their teaching philosophies. The motivation of MOOC instructors has been examined in a prior study [Doo20]. The majority of instructors' motivation to teach MOOCs is primarily related to intrinsic motivation. Some studies investigated the strategies and the challenges that MOOC instructors met when they designed a MOOC [Sar20]. Gitinabard et al. [Git22] designed a dashboard for instructors to monitor students' teamwork for team coding projects, and Akintunde et al. designed a dashboard [Aki21a] allowed instructors to get aware of students' different type of forum activities. Their works focused on displaying different types of students' activities on different platforms, such as discussion forums, GitHub, and MDH. Similar to their works, we designed an instructor dashboard to monitor students' activities in the discussion forum. Different from previous studies, instead of the number of activities, we focused on analyzing students' post content from the semantic aspect. Thus, we found there is little research on developing content-based, scaffold-based tools upon investigating instructors' teaching experience.

To address this challenge, we found prior researchers discussed how teachers use a dashboard designed with students' data to help students learn in a traditional classroom [Ale10; Xha16]. They designed the Intelligent Tutoring Systems (ITS) based on dashboard, named Luna, and investigate how a dashboard prototype designed for an ITS affect teachers' knowledge of their students. They argued that although ITS has been well studied, the reports from ITSs may not be designed to serve teachers' needs well. In their work, they discussed what student data is most helpful to teachers and how they use data to adjust and individualize instruction in a traditional class. They found that teachers generate data on students' concept mastery, misconceptions, and errors provided by an ITS. Teachers use this data to drive instruction on an individual and class level. They found that teachers mostly focused on student performance and the record of the individual student from the ITS. Then, they identified how teachers use data to drive instruction and help students. They found teachers want to know if the student is ready to:

- Move on to the next topic or build on current concepts. They want to know the status of students' learning to determine whether or not students learn a concept.
- Display the anonymous class report to students. In this case, teachers aim to support students' learning and progress by seeing how they were compared to the other students in the class.

Thus, to investigate instructors teaching experience and to develop a dashboard to support it, we aim to answer the following research question with this study: How do automated structures and content-based scaffolds support classroom orchestration and learning? To answer this question, we first recruited instructors to complete a text based survey. We then recruited them to take a video interview to provide more details of their learning experience and evaluate the designed dashboard from Chapter 7.

8.2 Literature Review

Qu et al. [Qu15] built a dashboard to visualize clickstream data for MOOCs. They showed the students' video clickstream actions at different time periods during the day and also included a social network built from the forum with students' grades. Leon et al. [Leó16] created a dashboard with students' weekly visualization of clickstream data and their social network. Cobos et al. [Cob16] designed a dashboard with student online interactions to allow instructors to learn students' participation in the forum and the web resources. However, none of these studies provided proper evaluation methods on their designed dashboard. Most of the them evaluated their dashboard from the usefulness aspect, but they didn't investigate instructors' workflow or what instructors need to support their teaching. As for instructor focused research on dashboard, Johnson et al. [Joh13] developed a instructor focused dashboard: InVis. InVis created a tree-like graph to represent the interaction network extracted from students' educational system usage data. It displayed the steps that students solve problems in the educational system. As a consequence, the interaction network based on students' problem solving steps enables instructors to make new insight and discoveries about student learning. On the other hand, Akintunde et al. [Aki21a] and Gitinabard et al. [Git22] developed two different visualization dashboards base on Piazza, a public discussion forum platform, for a blended course. Akintunde et al. developed a dashboard to analyze and visualize students' forum activities. It monitored students' help seeking behaviours by displaying students activities to instructors. Those activities defined as: not active, post followup, reply to followup, answer, and question. Gitinabard et al. developed a dashboard focused on monitoring students team work. This dashboard displayed how many activities does each team member made in different platforms, including GitHub, Piazza, And MyDigitalHand [Smi17].

In addition, Jivet et al. [Jiv18] conducted a survey of dashboard evaluations for educational practice. They found that very few dashboard evaluations take into account the educational concepts that were used as a theoretical foundation for their design. More specifically, they found out that evaluating a dashboard's acceptance, usefulness, and ease-of-use, as perceived by learners, is central in many of the analysed papers. "In almost all the cases, dashboard designers used feedback questionnaires and interviews for confirmation that learners are satisfied and find the visualizations useful". Among these works, the survey used to evaluate the dashboard asked people to rate the usability and usefulness of the tool. However, although the satisfaction of the tool is important, the primary focus of the dashboard should be whether the dashboard brings any benefit to instructors or learners.

In summary, previous studies designed the dashboard for instructors to monitor students in terms of their activities in different platforms/learning management systems. Unlike them, we focused on displaying student forum posts in the dashboard. We analyzed the semantics of discussion forum posts to understand what students were discussing and provided this information to instructors to monitor topics that students were discussing weekly.

8.3 Methods

8.3.1 Recruitment

In order to investigate MOOC instructors teaching experience, teaching philosophy, and evaluate the designed dashboard, we designed two rounds of interviews. For the first round interviews, we recruited MOOC instructors from all online public MOOC platform, such as Coursera and Edx. We reached out them individually via email attached with an online survey. For participants, the instructor should have experience teaching MOOCs and traditional classes. We set the criteria to ensure that the motivations and challenges that emerged from our interviews were unique artifacts of teaching MOOCs, as opposed to general struggles associated with prepping new courses or more generally teaching traditional university courses. In order to learn the instructors' teaching experience in general, we recruited instructors from our university including professors and graduate students pursuing doctorate degree that have served as instructors. We sent out surveys to 70 MOOC instructors within the Computer Science domain and 30 to all other domains, such as finance, literature, etc. We received 4 Computer Science MOOC instructor responses. For university instructors, we sent out surveys to 60 NCSU CSC faculty and 10 graduate student instructors who teach CSC summer courses. We receive 6 responses from the faculty and 4 from the graduate student instructors.

In order to learn their teach experience and teaching philosophy in details and evaluate our dashboard, we scheduled a second round interview among those who did the first round survey interview. We were able to complete 5 video interviews, 2 from faculty and 3 from graduate student instructors.

8.3.2 Participants Procedure

The participants consisted of the following steps:

- Completed a consent form.
- Completed a first round Google Form survey about their teaching experience and workflow. Questions are attached in the Appendix section.
- Provided their availability for the video interview.
- Answered questions about the details of their teaching experience and workflow based on their answer in the survey from Step 2.
- Provided feedback on the usefulness, effectiveness, and suggestions to the designed dashboard.

8.4 Results

8.4.1 MOOC instructors first round survey feedback

The results listed below are summarized from the 4 MOOC instructors we received responses from. The full content of the survey is attached in the appendix.

- *How you course organized*:
  Three of the instructors organized their course as typical MOOCs by different units including lecture videos, quizzes and longer assignments which were all automatically graded. One instructor misunderstood this question.

- *Is your course structure guided by a specific teaching philosophy, if so what is it?*:
  One instructors emphasized that he structured his course by setting more exercises to support the video lectures. He first provided a set of practice with provided answers related to the lecture videos and quizzes without provided answers. He also provided a forum for students and expected them to discuss course content. Another instructor structured the MOOC the same as the his traditional class, only divided into more weekly chunks. Two other instructors didn't respond to this question.

- *Are course activities strictly scheduled? Strictly ordered? Or are they self paced and self-ordered?*:
  Three instructors scheduled their courses activities to be completed strictly in order. One instructor scheduled the course by semi-self-order. There was a structure in the course, but the students don't have to follow it.

- *How many types of manually and automatically graded activities do you typically provide in a MOOC?*:
  Three instructors stated they will not manually grade activities themselves at all, and that most of the grading is automatic. One instructor grades one or two assignments via peer review and the automatically grades the rest.

- *How are students supported in the course?*:
  Three instructors don't support students with TA staff. One instructor said that his MOOCs were supported by either one full-time TA or by some part-time undergraduate students who hold small-group meetings with MOOC students. All of their students support their peers through the discussion forum.

- *How much time and effort do you (or your teaching assistants) spend grading or evaluating your MOOC students each week?*:
  Two instructors didn't spend anytime on grading or evaluating students. Two instructors said that they spend as little time as possible on grading.

- *How much time and effort do you (or your teaching assistants) spend replying students in your MOOC each week?*:
  They spend very little time interacting with students individually. Two instructors spend less than an hour and one spends less than 30 minutes.

- *Would you like to automatically monitor students' forum topic weekly?*:
  Two instructors would like to monitor students' weekly forum topics and one said he would maybe use the feature. One instructor didn't care about students' forum topic weekly at all.

- *Would you like to automatically monitor students' social connections at realtime?*:
  Three instructors said no to monitoring students' social connection at real time. One said maybe.

As a summary, MOOC instructors spend time preparing the course materials. Once they have setup the structure and the content of the course, they won't spend much effort to maintain the course. They won't actively interact with students and they will try their best to make all the work automatic.

8.4.2 University instructors Survey Case Study

For university instructors, we received 10 survey responds. The results listed below are summarized from their responds. The full content of the survey is attached in the appendix. The case studies below are from instructors who completed the survey but did not participate in the interview.

8.4.2.1 Case A

Professor A is a faculty member in the CSC department. He is an experienced professor whose taught graduate level Human-Computer Interaction (HCI) courses with about 100 students. His course had two teaching assistants and one grader. He used a Piazza forum to support students asking and answering questions with each others and TAs. This course meets two days a week. One day for lecturing and another day for lab. In the lab session, students formed a group of 5 or 6 and practiced methods that they learned in the lecture.

8.4.2.2 Case B

Professor B is a faculty member in the CSC department. He is an experienced professor who has taught under-graduate level Data Structure courses. His course had three sections, two in-person courses with about 120 students and one distance education section with about 30 students. He has had 5 teaching assistants and 1 undergraduate grader. He structured his course into 3 different units. Each units contains lecture topics, homework activities, and end with an exam. His course supported by the following tools: Moodle, Piazza, MyDigitalHand (an online office hour tool) [Smi17], Jenkins Continuous Integration Server, and GitHub. He encouraged and supported peer help seeking through his Piazza forum and he answered more than 90% of questions in the forum. To answer student questions, he spent 10-15 hours each week replying to student questions over email, office hours, and the Piazza forum. Each of his TA spent 3 hours in office hours and at most 1 hour on Piazza. He only supported and approved Piazza forum to discuss course assignments because of widespread cheating/aiding & abetting concerns that had increased as course enrollments increase. He was aware some students used external Discord, Slack, etc. applications to communicate about course content. However, students were not allowed to discuss course assignments on any other platform than Piazza. He would like to if the same most common questions each week were posted by the same group of students. Because, as he said, "Very often, questions on Piazza make it clear some students have not read assignment instructions or have not taken time to review lecture materials where the exact answers to their questions are fully addressed in the slides or in the lecture recordings". He would also like to monitor students' collaborative interactions by knowing the statistics about the discussions vs. timeline, i.e. how far ahead of deadlines are students discussing course topics.

8.4.2.3 Case C

Professor C is a faculty member in the CSC department. He is an experienced professor whose taught graduate level Computer and Network Security. He had one graduate teaching assistant and one grader. He used Piazza to support students and mentioned it worked well. He also found students weren't terribly active on the forum, so it was managed only by himself. He and his TA each spent less than 2 hours to reply to students questions through Piazza and direct emails. He may or may not like to automatically monitor students weekly questions because he thought Piazza works well in this function. He also may or may not want to monitor students' collaborative interactions because he felt like it could be invasive to students' privacy.

8.4.2.4 Case D

Instructor D is a graduate student instructor who taught Discrete Mathematics for Computer Scientists. His course was during a summer semester, had about 20 students, and did not have a teaching assistant. To interact with students, he held weekly office hours, answered questions in the Piazza forum, and paused a lot during his lecture to ensure no students had concerns or questions. He spent less than one hour each week answering students' questions because he stated students don't ask many questions in his class. He may or may not want to automatically monitor students questions and thought Piazza met his needs to answer and monitor students learning status.

8.4.2.5 Case E

Professor E is a faculty member in the CSC department. He is an experienced professor whose taught online Discrete Mathematics for Computer Scientists during summer semesters with about 40 students and one teaching assistant. They used Moodle platform to answer students questions and support peer discussion. They spent 1-2 hours replying to students questions each week in addition to office hours. He mentioned email is the most used method by students instead of class forum. He would not like to monitor students common questions because he encourages students to use the forum to ask questions so that everyone can see it. He also emphasized that the volume of questions isn't overwhelming, so that the forum is enough for him to support students' learning. He would like to monitor students' collaborative interactions by whether or not students do collaborate learning.

8.4.3 University instructors Survey and Interview Case Study

We interviewed five instructors, who also responded to the Google Form survey, to investigate their teaching experience in details and their feedback on the designed dashboard.

8.4.3.1 Case F

Instructor F was a graduate student instructor who taught Intro to Computing in Java. Her course was during a summer semester through online lecture with about 30 students and one teaching assistant. She supported and encouraged students to support their peers' learning through Piazza. She and her TA spent a lot of time replying students' questions. They would also answer questions during in-class lab session, held weekly office hours, and took care of emails and Piazza posts related to programming and debugging everyday. Although they encouraged students to post questions on the forum, many students still preferred emails. She would like to monitor students' common questions before each lecture, so that she could address them immediately.

As for the designed dashboard, she thought the dashboard of hot topic words is "definitely helpful to instructors". When it came to the common questions and course materials, she thought this would be more helpful to large scale instructors. Since she only had about 30 students, it would not be a very efficient tool for her to use. In addition, she mentioned as she taught a intro level of programming course. Most of the questions she received were debugging requests, so the material recommendation would probably not help her solve students' questions. As for suggestions, she would like to link the tool with Piazza directly for efficiency.

8.4.3.2 Case G

Instructor G was a graduate student instructor who taught Software Development Fundamentals. His course was during a summer semester through online lecture with about 60 students and 3 teaching assistants. His course was based heavily around active learning and team-based learning. He used Piazza for students to discuss with their peers. His students were not allowed to share code with each other, but he encouraged them discuss concepts because concepts can be more easily described in a general sense and answered by other students. Each week, he spent about 6-8 hours replying to student questions and each of his TAs spent about 2-3 hours. Beside Piazza, he also found out that his students used Discord to discuss course content. He would like to monitor students common questions and collaborative interactions each week. With common questions, he can prepare a Q&A to answer similar questions from students. He also mentioned it would be very useful to better understand how students collaborate and how much each team member contributes to the team.

G liked the idea of the dashboard and believed that it could support his teaching. However, he would like to have more information and a better interface of it. For example, he said beside the summarization of the selected thread, it would be better to have other replies under it. Showing the frequency of hot topics and common questions would also help him understand more about students' learning status.

8.4.3.3 Case H

Professor H is a faculty member in the CSC department. She is an experienced professor whose taught Software Development Fundamentals. Her course was a large scale course with about 400 students and 15-20 teaching assistants. TAs were expected to answer questions on the Piazza forum and would discuss questions they can't answer in another message board group. They spent about 2 hours replying to students question each week through Piazza and emails. They also supported the Discord platform to make announcements by a TA but it was not encouraged by the instructor. She would like to monitor students common questions and whether students work in an appropriate way in team-projects. She would also like to monitor how students use the message board, office hours, and how these features tie to student course performance.

Professor H mentioned that the hot topics and common questions could helpful. However, she identified some drawbacks to the dashboard. First, the user interface could be problematic when there are a lots of threads related to the hot topics. Second, summarizing threads from the header post is dangerous, some students wrote poor header posts during her observation from her teaching. She defined poor header posts in this situation as those header posts that do not match what the student wants to ask. Finally, she thought the dashboard had too much text for her to use. It would be better to extract more tags from the selected thread and then those tags could save more time and make it more clear to instructors what happened in the student discussion forum.

8.4.3.4 Case I

Instructor I was a graduate student instructor who taught Intro to Artificial Intelligence. His course had about 60-90 students and 2 teaching assistants. He supported peers help seeking through Piazza. Besides office hour, he and his TAs would spend about 1 hour replying to students questions. He mentioned that the most popular questions are typically clarifications on assignments/questions or logistics about course set up. On the other hand, as deadlines approached, debugging help was another common question type. His course also had an unofficial Discord channel where his TAs were identified with roles; however this channel was not meant to be used for asking teaching staff questions. He said monitoring students question may or may not help his teaching because his students often asked questions that are expecting immediate response or were heavily contextual to their work. For example, students would often post pictures of program errors. He would like to monitor students to know what materials students are struggling with.

As for the dashboard, he thought the hot topic words and most common questions would be useful. He suggested a frequency for each topic and questions would help instructors to know what students struggling and whether instructors need to answer the question immediately.

8.4.3.5 Case J

Professor J is a faculty member in the CSC department. He is an experienced professor whose taught graduate level Computer Networks with a small classroom size about 30 students and 0.5 to 1 teaching assistant. He mentioned his teaching philosophy is "fairness-first". For example, some students worked on their assignments close to the deadline and some did not. In this case, he tried to design the assignment and course structure to be "fair" for both working styles. He and his TA spent 1-2 hours each week replying to students' questions mostly through Piazza and a few by direct emails. He also emphasized that very few of students came in to the office hour and that students were more likely to use Piazza to seek help. He would like to monitor topics and numbers of questions. He also would like to monitor students' collaborative interactions from the demographics of the teach, such as topics, group size, and time length.

As for the dashboard, he said the hot topics could help him learn more on students' learning status from the forum. However, he may not need this tool because he has a small size of classroom and student didn't have a lot of questions during his course.

8.4.4 Summarization

As a summary, university instructors design their courses within a similar course structure with MOOCs, i.e., divide classes into lectures, videos, and quizzes/assignments. University instructors spend much longer time interacting with students by answering their questions or providing feedback to their submissions. They also care about inappropriate collaborate behaviours during students collaborate learning.

As a general conclusion, university instructors considered interaction with students as one of the most important tasks for them. The effort they spend to helping students was strongly dependent on the quality of their TAs. In addition, most of them preferred students to first seek help from their peers or from other resources such as text book and internet before seeking help from the instructor. They believe that kind of help seeking behaviour will make the office hour or email more effective. However, some of the instructors pointed out that for the introductory university courses, it is also a task for instructors to teach students on how to efficiency ask questions.

We interviewed university instructors to provide feedback on the dashboard we designed in Chapter 7 from three different aspect: whether the information is useful, how efficiency of the dashboard, and what can be improved. Most of the instructors believe this forum discussion monitor tool could provide an insight of the students' forum behaviors. However, when it comes to the efficiency and whether it can save their time diving into the forum, it depends on the population of the course. Some instructors pointed that they have a small size of classroom and it probably wouldn't take them too much time to go through their forum every week. However, population of the classroom is not the only element impact instructors workload of using the discussion forum. The level of the course is another factor. Some instructors(instructor F, G, H, and I) mentioned during the video interview that for low level CSC courses they taught, students were more likely to ask more questions in the public forums on the course concepts and assignments. Besides, Instructor F, G, and I also mentioned that they will encourage low level CSC courses' students ask more questions directly in the discussion forum. Thus, the CSC course level is also considered a factor in whether instructors would like to use the designed dashboard to support their teaching. In addition, instructors also mentioned that they would like to use this tool if there are more human computer interaction components. For example, instructor would like to see the frequency of each topic word and new questions to decide whether they need to provide support immediately.

Chapter 9

Conclusions

My goal in this work is to provide methods to organize MOOC discussion forums and provide support to instructors based upon research findings. Current studies are insufficient because they lack details discussion of the students discussion forum behaviours, lack of anaslysis of instructor-based experience, and lack of applying content-based scaffolds to support instructors teaching experience. To address the shortcomings of the current approach, this dissertation investigating the following research questions:

- RQ1: How can we model the communicative activities of MOOC students based on their forum posts?
  - RQ1.1: How do students build their communication over time based on their forum reply relationship?
  - RQ1.2: How do we understand students forum post by categorizing the post content?
  - RQ1.3: How do we analyze student online activities transaction combined their click-stream and forum activities?
- RQ2: How can instructors monitor students' learning status based on their use of the forum? And can we use our knowledge of structured forum to help instructor make initial interventions and improve students learning experience and students get answers more efficiently?
  - By efficiently partitioning requests for help from other contexts.
  - By clustering similar questions to recognize duplicate and non-answered topics.
  - By aligning questions to related course materials.
- RQ3: What are instructors' goals in monitoring student interactions? How do they evaluate adaptive scaffolds? What features do they want in an adaptive scaffold for content monitoring?

9.1 RQ 1:How can we model the communicative activities of MOOC students based on their forum posts?

To address RQ1, I first analyzed how students form their social connections in the public discussion forums over time. I found out that through the entire time period of MOOCs, good performance students are more likely to connect with good performance students and bed performance students are more likely to connect with bad performance students. An limitation of this study is we applied the statistical analysis to generate the conclusions. However, this correlation could caused by zero-grade students communicating with one another rather than communicating with non-zero. The future work will need to examine the proportion of communication within a grade level. Then, to understand students forum post content, I developed a coding schema specific to MOOCs and evaluate it by reliability analysis. However, imbalanced data is the biggest problem for automatic classification. Widely used methodologies, such as under sampling, oversampling, and SMOTE do not work in this circumstance. Thus, I developed a cascade structure to address the imbalanced data problem. The main idea of the cascade structure is among all classes, first use binary classifier to classify the most population class, and iterate the step until all data point has been classified. This methodology not only works for my dissertation work, but also works for other educational data mining projects [Git20]. In addition, to improve the performance of the automatic classifier, I introduced parent-child relation features. The idea is based on the assumption that in a public discussion forum, the neighbor post must related. As a result, this feature improved the F1 score of the classifier. Finally, I analyzed the sequential behaviours by combining students' clickstream log actions and their discussion forum post actions. I found out that good performance students always make more transactions between the course lecture/assignments and the forum. These finding could help researchers and teaching staff understand more about MOOC students' behaviour patterns.

9.2 RQ 2:How can instructors monitor students' learning status based on their use of the forum?

To investigate the research question 2, this work applied the widely used methodologies in the text mining domain to monitor MOOC discussion forums. First, I applied the Latent Dirichlet Allocation algorithm to monitor students' weekly hot topics. Then I developed a dashboard for instructors to monitor students' forums activities based on text summarization and similarity techniques. Eventually, the dashboard displays hot topic keywords of different classes of posts topics, the summarization of the thread information related to selected topic words, and the most recent new questions with the recommended similar previous threads and related course materials title. These extracted information from the students' discussion forums reflect students' social learning behaviours and help-seeking behaviours. Finally, I evaluated the dashboard by interviewing five university instructors in the form of video. Based on their feedback, they agreed that the designed content-based dashboard could help them save time to monitor students forum activities for their courses. For those instructors who have many students (larger than 40 students) in their classroom, they emphasized that the dashboard could help them understand students' learning status from the forum efficiently. However, they would like to use it if it can provide more information and human interaction modules, such as connecting this tool directly with Piazza and adding more tags to the summary of threads.

9.3 RQ 3:What are instructors' goals in monitoring student interactions? How do they evaluate adaptive scaffolds? What features do they want in an adaptive scaffold for content monitoring?

To examine research question 3, I sent out a survey to MOOC instructors and university instructors to investigate their teaching experience and teaching philosophy. I investigated whether the instructors would like to monitor students and what information they would like to receive from the students. In summary, although the instructors that I interviewed stated that social collaboration is one of the most crucial elements in MOOC learning, they also made clear that they spent much less time interacting with their students, and they are more willing to automate their tasks. On the other hand, CSC university instructors that I interviewed stated that they consider interaction with students to be the most important part of their teaching. They would like to spend more time responding to students' questions either through the email or the public discussion forums. Thus, the content-based tool which focuses on students' questions, and interactions was more attractive to them. The information collected from the survey lets researchers know what instructors need to monitor students learning status and thus, improve instructors' teaching experience. We have already investigated how many time did each instructors spent in the discussion forum as a quantitative feature to measure the size of discussion forums. However, to address these questions more concretely and to draw stronger conclusions we would need to interview a larger group and add follow-up questions about how many messages of posts they will need to reply in average each week.

9.4 Limitation & Future Work

9.4.1 RQ 1.1:How do students build their communication over time based on their forum reply relationship?

We break down the MOOC social network based on a two-week chunk to analyze how students form their forum connections. To build students' social networks, we assume that MOOC students will only use the discussion forum to communicate with other people. Although the discussion forum is the only official venue for them to share their knowledge, based on our observation of the forum discussion, students will introduce themselves and establish chat groups on other social media such as Facebook. Thus, there is a possibility that MOOC students make friends and connections outside the discussion forum. An area for future work will be to investigate MOOC students' communication habits from the survey, so researchers and instructors can learn the importance of the discussion forum in MOOC students' social communication. Another limitation of this study is we applied the statistical analysis to analyze the correlation between students' performance and their best friends' (one-step neighbors in the social network) performance. However, this correlation could caused by zero-grade students communicating with one another rather than communicating with non-zero. The future work will need to examine the proportion of communication within a grade level.

9.4.2 RQ 1.2:How do we understand students forum post by categorizing the post content?

One limitation of this work is that we create the coding schema not only based on the previous work about how different content posts impact students' learning but also on our observation of the dataset. Thus, the generalization of the coding schema is a limitation of this work. An area for future work will be to offer this coding schema to other MOOCs so that we can verify the generalization of this coding schema or revise the coding schema. On the other hand, MOOCs are open to worldwide learners who have internet access, and other languages instead of English have the potential to occur more in the discussion forum. Thus the interpretation of non-English posts could be a problem in understanding all students' learning status.

9.4.3 RQ 1.3:How do we analyze student online activities transaction combined their clickstream and forum activities?

We utilize a sequential relationship between videos and forum actions to investigate students' online learning behaviours, but the hierarchies of the MOOC syllabus are not considered in this paper. This information could help instructors to understand what students do on the website and what they post in the forum. Future work will identify students' online behaviour patterns based on different periods of the syllabus, for example, what they do around the deadline of assignments. Akintunde et al. [Aki21b] analyzed students' forum activities near assignments deadline. They found out that students are often deadline-driven and begin work until near deadlines. This information could help instructors to make interventions to encourage MOOC students' engagement in the course and improve their learning outcomes. On the other hand, we cannot collect the private data of the chat room from the Edx platform. Students may chat with other people while working on the lecture and assignments. Thus, another future work will be to survey students to explore what and when they discuss using the private chat room.

9.4.4 Research Question 2

We focus on monitoring the dashboard for instructors to save their time and improve their teaching experience. In terms of internal validity, though we reached the accuracy of 0.8 between two human coders and the topic modeling algorithm output, the topics that topic modeling generated still have a possibility that was not capturing the hot topics that need attention from the students side. Thus, future work will need to validate the topic modeling algorithm results from instructors and students to confirm the generated topics are hot topics discussed by students.

In terms of external validity, we evaluated the dashboard design based on a few computer science instructors so that our conclusion has a bias that cannot be generalized to the entire computer science instructors population. In addition, we designed the dashboard based on MOOC data and generalized it to university blended computer science courses. However, there is a difference between these two types of course. For example, the discussion forum is the only venue for MOOC students to communicate, but for blended courses, students have a chance to communicate in the classroom face to face.

In terms of construct validity, the two coders who validated the output of the topic modeling algorithm came from the similar education background and both of whom have not take part in the annotated course. However, during the annotation procedure, we assumed that both coders understand the course deeply and are able to figure out hot topics from students correctly. Thus, future work will need to hire an expert instructor to validate the human annotated hot topics results.

9.4.5 Research Question 3

We applied quantitative research to analyze the instructors' survey results. First, quantitative research is usually conducted on a section of the target population, not the whole population. The outcome of this research is then generalized as the view of the entire population. In this study, we interviewed just a few computer science instructors, so the findings may not be generalized. Such views generated from the interviewees might be biased or insincere when we generalize the views to the entire population.

Second, in this work is that we didn't focus on the gender of the interviewee. Prior studies [She11; Fan19] found that the gender of instructors impacted student evaluation. Thus, future work will need to submit an edition in Institutional Review Boards documents on analyzing the impact of the gender of the interviewee. And then apply quantitative analysis to the survey results from the aspect of gender.

Finally, we discussed how class size/scale affects the instructors' attitude on using the designed dashboard. However, during the interview, we did not consider the differences between the class size and if the instructor feels that they can manage the discussion forum on their own. The number of posts may not only be influenced by the student population but also by the class level, the instructor's teaching style, etc. Thus, future work will need to collect the number of posts in the discussion forum each week to examine the factor related to instructors' need for the dashboard to monitor students' forum activities.

Chapter 10

Appendix

10.1 First Round Survey for MOOC Instructions

- How many MOOCs have you taught in the past 2 years and on what topics?
- How is your course organized?
- Is your course structure guided by a specific teaching philosophy, if so what is it? For example, ideally, how would you prefer students consume and engage the learning material for any given module? If it's couldn't be described briefly in the answer box, would you like to schedule a meeting with us?
- Are course activities strictly scheduled? Strictly ordered? Or are they self paced and self-ordered?
- How many types of manually and automatically graded activities do you typically provide in a MOOC?
- How are students supported in the course?
- How long have you been teaching MOOC? How many different topics have you taught?
- What is your non-MOOC teaching experience?
- What technology do you use to support your MOOC courses? And what do you use to support other non-MOOC courses you teach?
- How much time and effort do you (or your teaching assistants) spend on preparing materials for your MOOC each week?
- How much time and effort do you (or your teaching assistants) spend grading or evaluating your MOOC students each week?

- How much time and effort do you (or your teaching assistants) spend replying students in your MOOC each week?
- What features of students' discussions, habits, or interactions would you like to monitor and how would you use that information?
- Do your students use other platforms (e.g. meetups, Slack, etc.) to communicate about course content? If so is this something you have encouraged or discouraged?

10.2 First Round Survey for University Instructions

- What is the name/number of this course?
- How did you structure the content in this class?
- Would you be willing to share your syllabus and schedule with us?
- Was your class design based upon a specific teaching philosophy or instructional model? If so, can you describe it or would you be willing to discuss it in a meeting?
- How do you separate the content of your most recent course topics? By assignment? By Module? As part of an assignment?
- How do you and the teaching staff support students' learning in your course?
- How many TAs do you usually have supporting your class?
- Do you use peer support via forums? If so, how is that managed?
- What technologies & platforms do you use to support your class?
- In general, how much time would you say that you and your TAs spend preparing materials for the course each week?
- How much time would you say that you and your TAs spend grading or evaluating your students each week?
- In general how much time would you say you and your TAs spend replying to student questions each week?
- What platforms / venues do students typically use for questions in your class and which ones are the most popular?
- Do your students use other platforms (e.g. meetups, Slack, etc.) to communicate about course content that you are aware of? If so is this something you have encouraged or discouraged?

- Do you think it would be beneficial to automatically monitor and summarize students' class questions each week?
- If so, what would you like to know about them?
- Do you think it would be beneficial to automatically monitor and summarize students' collaborative interactions each week?
- If so, what would you like to know about them?
- Is there any other aspect of students' discussion habits, or interactions that you would like automatically monitor?

10.3 Second Round Interview for University Instructions

Questions Part 1:

- In your opinion how should a course be structured ideally in order to be useful, effective, and efficient?
- What motivated you to design and offer your course? And what analysis have you done of it?
- What do you see as your primary role in your course? Are you more invested in the overall organization, lectures, assignments, grading, or direct support of students?
- How important is instructor feedback and engagement with students to the course experience? How much time do you spend on that and how much effect do you think it has?
- How do you view the students' engagement with the course? Should they be focused on seeking help from you, from their peers, or from other sources?
- How would you rate the interactions that you have had with students in your course? Do you find the experience informative, manageable, or problematic? And why?

Questions Part 2:

- How do you think the efficiency of this dashboard will help you learn more about students' forum discussions, please rate from 1 to 10.
- How do you think the useful information provided by this dashboard will help you reduce the effort of answering students' questions, please rate from 1 to 10.

BIBLIOGRAPHY

[Ada08] Adamic, L. A. et al. "Knowledge sharing and yahoo answers: everyone knows something". *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 665–674.

[Ade16] Adel, S. M. R. et al. "A qualitative study of politeness strategies used by Iranian EFL learners in a class blog". *Iranian Journal of Language Teaching Research* **4**.1 (2016), pp. 47–62.

[Agr15] Agrawal, A. et al. "YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips" (2015).

[Agu09] Aguilar, D. A. G. et al. "Semantic spiral timelines used as support for e-learning." *J. Univers. Comput. Sci.* **15**.7 (2009), pp. 1526–1545.

[Aki] Akintunde, R. O. et al. *Exploring class level forum usage for instructor insights*.

[Aki21a] Akintunde, R. O. et al. "PEDI-Piazza Explorer Dashboard for Intervention". *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE. 2021, pp. 1–4.

[Aki21b] Akintunde, R. O. *Learning Analytics for Data-Informed Interventions in Blended Classrooms*. North Carolina State University, 2021.

[Alb14] Albitar, S. et al. "An effective TF/IDF-based text-to-text semantic similarity measure for text classification". *International Conference on Web Information Systems Engineering*. Springer. 2014, pp. 105–114.

[Ale10] Aleven, V. et al. "Developing a teacher dashboard for use with intelligent tutoring systems". *technology* **34** (2010), pp. 44–50.

[Alj13] Aljohani, N. R. & Davis, H. C. "Learning analytics and formative assessment to provide immediate detailed feedback using a student centered mobile dashboard". *2013 Seventh international conference on next generation mobile apps, services and technologies*. IEEE. 2013, pp. 262–267.

[Ana15] Anaya, A. R. et al. "An Approach of Collaboration Analytics in MOOCs Using Social Network Analysis and Influence Diagram." *EDM*. 2015, pp. 492–495.

[And14] Anderson, A. et al. "Engaging with massive online courses". *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 687–698.

[Arc99] Archer, W. et al. "Adopting disruptive technologies in traditional universities: Continuing education as an incubator for innovation". *Canadian Journal of University Continuing Education* **25**.1 (1999).

[Ata16a] Atapattu, T. & Falkner, K. "A Framework for Topic Generation and Labeling from MOOC Discussions". *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*. L@S '16. Edinburgh, Scotland, UK: ACM, 2016, pp. 201–204.

[Ata16b] Atapattu, T. et al. “Topic-wise Classification of MOOC Discussions: A Visual Analytics Approach”. *EDM*. International Educational Data Mining Society (IEDMS), 2016, pp. 276–281.

[Bal13] Balakrishnan, G. & Coetzee, D. “Predicting student retention in massive open online courses using hidden markov models”. *Electrical Engineering and Computer Sciences University of California at Berkeley* **53** (2013), pp. 57–58.

[Bal00] Baldi, P. et al. “Assessing the accuracy of prediction algorithms for classification: an overview”. *Bioinformatics* **16**.5 (2000), pp. 412–424.

[Beh16] Beheshitha, S. S. et al. “The role of achievement goal orientations when studying effect of learning analytics visualizations”. *Proceedings of the sixth international conference on learning analytics & knowledge*. 2016, pp. 54–63.

[Ber08] Berkovsky, S. et al. “Mediation of user models for enhanced personalization in recommender systems”. *User Modeling and User-Adapted Interaction* **18**.3 (2008), pp. 245–286.

[Ber88] Berry, K. J. & Mielke Jr, P. W. “A generalization of Cohen’s kappa agreement measure to interval measurement and multiple raters”. *Educational and Psychological Measurement* **48**.4 (1988), pp. 921–933.

[Ble03] Blei, D. M. et al. “Latent Dirichlet Allocation”. *Journal of Machine Learning Research* **3** (2003), pp. 993–1022.

[Bon18] Bonk, C. J. et al. “Pushing toward a more personalized MOOC: Exploring instructor selected activities, resources, and technologies for MOOC design and implementation”. *International Review of Research in Open and Distributed Learning* **19**.4 (2018).

[Bor17] Boroujeni, M. S. et al. “Dynamics of MOOC discussion forums”. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 2017, pp. 128–137.

[Boy15] Boyer, S. & Veeramachaneni, K. “Transfer learning for predictive models in massive open online courses”. *International conference on artificial intelligence in education*. Springer. 2015, pp. 54–63.

[Bri15] Brinton, C. G. & Chiang, M. “MOOC performance prediction via clickstream data and social learning networks”. *2015 IEEE conference on computer communications (INFOCOM)*. IEEE. 2015, pp. 2299–2307.

[Bri16] Brinton, C. G. et al. “Social learning networks: Efficiency optimization for MOOC forums”. *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications* (2016), pp. 1–9.

[Bro15a] Brown, R. et al. “Communities of Performance & Communities of Preference.” *EDM (Workshops)*. 2015.

[Bro15b] Brown, R. et al. "Good Communities and Bad Communities: Does membership affect performance?" *EDM (Workshops)*. 2015.

[Cad10] Cade, W. L. et al. "An exploration of off topic conversation". *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010, pp. 669–672.

[Caf99] Caffarella, R. S. & Merriam, S. B. "Perspectives on adult learning: Framing our research". *40th Annual Adult Education Research Conference Proceedings*. Northern Illinois University DeKalb, IL. 1999.

[Cal10] Calvani, A. et al. "Visualizing and monitoring effective interactions in online collaborative groups". *British Journal of Educational Technology* **41**.2 (2010), pp. 213–226.

[Che16] Chen, Y. et al. "Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals". *Proceedings of the sixth international conference on learning analytics & knowledge*. 2016, pp. 1–5.

[Che11] Chenghui, H. et al. "A text similarity measurement combining word semantic information with TF-IDF method". *Chinese journal of computers* **34**.5 (2011), pp. 856–864.

[Chi14] Chi, M. T. & Wylie, R. "The ICAP framework: Linking cognitive engagement to active learning outcomes". *Educational Psychologist* **49**.4 (2014), pp. 219–243.

[Cho14] Choo, E. et al. "Revealing and incorporating implicit communities to improve recommender systems". *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM. 2014, pp. 489–506.

[Chr16] Christian, H. et al. "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)". *ComTech: Computer, Mathematics and Engineering Applications* **7**.4 (2016), pp. 285–294.

[Cob16] Cobos, R. et al. "Open-DLAs: An open dashboard for learning analytics". *Proceedings of the third (2016) ACM conference on learning@ scale*. 2016, pp. 265–268.

[Coe14] Coetzee, D. et al. "Should your MOOC forum use a reputation system?" *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, pp. 1176–1187.

[Coh94] Cohen, E. G. "Restructuring the classroom: Conditions for productive small groups". *Review of educational research* **64**.1 (1994), pp. 1–35.

[Cor95] Cortes, C. & Vapnik, V. "Support Vector Networks". *Machine Learning* **20** (1995), pp. 273–297.

[Cro15] Crossley, S. et al. "Language to Completion: Success in an Educational Data Mining Massive Open Online Class." *International Educational Data Mining Society* (2015).

[Cro16] Crossley, S. et al. “Combining click-stream data with NLP tools to better understand MOOC completion”. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM. 2016, pp. 6–14.

[Cui15] Cui, Y. & Wise, A. F. “Identifying content-related threads in MOOC discussion forums”. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. 2015, pp. 299–303.

[Cus15] Cusick, M. B. et al. “Human Generated Topics: A Gold Standard for Automated Topic Evaluation.” (2015).

[Dah14] Dahlstrom, E. & Bichsel, J. “ECAR Study of Undergraduate Students and Information Technology, 2014.” *Educause* (2014).

[Dal02] Dalgaard, P. *Introductory Statistics with R*. Springer Verlag New York Inc., 2002.

[Daw10] Dawson, S. “’Seeing’the learning community: An exploration of the development of a resource for monitoring online student networking”. *British Journal of Educational Technology* **41**.5 (2010), pp. 736–752.

[Des13] Desmarais, M. & Lemieux, F. “Clustering and visualizing study state sequences”. *Educational Data Mining 2013*. 2013.

[Doo20] Doo, M. Y. et al. “MOOC instructor motivation and career development”. *Distance Education* **41**.1 (2020), pp. 26–47.

[Eas18] Easterday, M. W. et al. “The logic of design research”. *Learning: Research and Practice* **4**.2 (2018), pp. 131–160.

[Eck12] Eckles, J. E. & Stradley, E. G. “A social network analysis of student retention using archival data”. *Social Psychology of Education* **15**.2 (2012), pp. 165–180.

[EC13] Ezen-Can, A. & Boyer, K. E. “Unsupervised classification of student dialogue acts with query-likelihood clustering”. *Educational Data Mining 2013*. 2013.

[EC15] Ezen-Can, A. et al. “Unsupervised Modeling for Understanding MOOC Discussion Forums: A Learning Analytics Approach”. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. LAK ’15. Poughkeepsie, New York: ACM, 2015, pp. 146–150.

[Fan19] Fan, Y. et al. “Gender and cultural bias in student evaluations: Why representation matters”. *PloS one* **14**.2 (2019), e0209749.

[Fer13] Ferguson, R. et al. “An evaluation of learning analytics to identify exploratory dialogue in online discussions”. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. ACM. 2013, pp. 85–93.

[Fir12] Fire, M. et al. “Predicting student exam’s scores by analyzing social network dataAnderson”. *International Conference on Active Media Technology*. Springer. 2012, pp. 584–595.

[Gau08] Gaudioso, E. et al. "Enhancing e-learning through teacher support: two experiences". *IEEE Transactions on Education* **52**.1 (2008), pp. 109–115.

[Git18] Gitinabard, N. et al. "Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features". *arXiv preprint arXiv:1809.00052* (2018).

[Git19] Gitinabard, N. et al. "What will you do next? A sequence analysis on the student transitions between online platforms in blended courses". *arXiv preprint arXiv:1905.00928* (2019).

[Git20] Gitinabard, N. et al. "Student Teamwork on Programming Projects: What can GitHub logs show us?" *arXiv preprint arXiv:2008.11262* (2020).

[Git22] Gitinabard, N. et al. "Designing a Dashboard for Student Teamwork Analysis". *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1*. 2022, pp. 446–452.

[Gol96] Goldberg, M. W. "Student participation and progress tracking for web-based courses using WebCT". *Proceedings of the Second International NA WEB Conference*. 1996, pp. 5–8.

[Gol07] Golder, S. A. et al. "Rhythms of social interaction: Messaging within a massive online network". *Communities and technologies 2007* (2007), pp. 41–66.

[Gru16] Gruzd, A. et al. "Analyzing social media and learning through content and social network analysis: A faceted methodological approach". *Journal of Learning Analytics* **3**.3 (2016), pp. 46–71.

[Hag18] Hagen, L. "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?" *Information Processing & Management* **54**.6 (2018), pp. 1292–1307.

[Hal14a] Halawa, S. et al. "Dropout prediction in MOOCs using learner activity features". *Experiences and best practices in and around MOOCs* **7** (2014), pp. 3–12.

[Hal14b] Halawa, S. et al. "Dropout prediction in MOOCs using learner activity features". *Proceedings of the second European MOOC stakeholder summit* **37**.1 (2014), pp. 58–65.

[Har99] Hardless, C. & Nulden, U. "Visualizing learning activities to support tutors". *CHI'99 extended abstracts on Human factors in computing systems*. 1999, pp. 312–313.

[Har04] Hardy, J. et al. "e-learner tracking: Tools for discovering learner behavior". *The IASTED International Conference on Web-base Education*. 2004, pp. 458–463.

[He15] He, J. et al. "Identifying at-risk students in massive open online courses". *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[Hin02] Hintze, J. M. et al. "Best practices in the systematic direct observation of student behavior". *Best practices in school psychology* **4** (2002), pp. 993–1006.

[Hos05] Hoskins, S. L. & Van Hooff, J. C. "Motivation and ability: which students use online learning and what influence does it have on their achievement?" *British journal of educational technology* **36**.2 (2005), pp. 177–192.

[HI17] Houston II, S. L. et al. "Pass the idea please: The relationship between network position, direct engagement, and course performance in MOOCs". *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. ACM. 2017, pp. 295–298.

[Hua14] Huang, J. et al. "Superposter behavior in MOOC forums". *Proceedings of the first ACM conference on Learning@ scale conference*. ACM. 2014, pp. 117–126.

[Ind10] Indurkhya, N. & Damerau, F. J. *Handbook of natural language processing*. Chapman and Hall/CRC, 2010.

[Ins16] Insa, D. et al. "Where You Sit Matters How Classroom Seating Might Affect Marks". *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. ACM. 2016, pp. 212–217.

[Ito13] Ito, M. et al. *Connected learning: An agenda for research and design*. Digital Media and Learning Research Hub, 2013.

[Jia14] Jiang, S. et al. "Social positioning and performance in MOOCs". *Workshop on Graph-Based Educational Data Mining*. Vol. 14. 2014.

[Jiv18] Jivet, I. et al. "License to evaluate: Preparing learning analytics dashboards for educational practice". *Proceedings of the 8th international conference on learning analytics and knowledge*. 2018, pp. 31–40.

[Joh06] Johnson, D. W. et al. "Improving computer literacy of business management majors: A case study". *Journal of Information Technology Education: Research* **5**.1 (2006), pp. 77–94.

[Joh13] Johnson, M. et al. "Invis: An interactive visualization tool for exploring interaction networks". *Educational Data Mining 2013*. Citeseer. 2013.

[Kay13] Kay, J. et al. "MOOCs: So Many Learners, So Much Potential..." *IEEE Intelligent systems* **28**.3 (2013), pp. 70–77.

[Kha19] Khan, R. et al. "Extractive based text summarization using k-means and tf-idf". *International Journal of Information Engineering and Electronic Business* **11**.3 (2019), p. 33.

[Kin12] Kinnebrew, J. S. & Biswas, G. "Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution." *International Educational Data Mining Society* (2012).

[Kiz13] Kizilcec, R. F. et al. “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses”. *Proceedings of the third international conference on learning analytics and knowledge*. ACM. 2013, pp. 170–179.

[Klo14] Kloft, M. et al. “Predicting MOOC dropout over weeks using machine learning methods”. *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. 2014, pp. 60–65.

[Kov14] Kovanovic, V. et al. “What is the source of social capital? The association between social network position and social presence in communities of inquiry” (2014).

[Kov15] Kovanovic, V. et al. “Penetrating the black box of time-on-task estimation”. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, LAK ’15, Poughkeepsie, NY, USA, March 16-20, 2015*. 2015, pp. 184–193.

[Kra14] Kramer, A. D. et al. “Experimental evidence of massive-scale emotional contagion through social networks”. *Proceedings of the National Academy of Sciences* **111**.24 (2014), pp. 8788–8790.

[Kri18] Krippendorff, K. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[Kru52] Kruskal, W. H. & Wallis, W. A. “Use of ranks in one-criterion variance analysis”. *Journal of the American statistical Association* **47**.260 (1952), pp. 583–621.

[Kul18] Kulkarni, M. & Boyer, K. “Toward Data-Driven Tutorial Question Answering with Deep Learning Conversational Models”. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 273–283.

[Lal04] Laland, K. N. “Social learning strategies”. *Animal Learning & Behavior* **32**.1 (2004), pp. 4–14.

[Lee99] Lee, D. D. & Seung, H. S. “Learning the parts of objects by non-negative matrix factorization”. *Nature* **401**.6755 (1999), pp. 788–791.

[Leh10] Lehman, B. et al. “Off topic conversation in expert tutoring: Waste of time or learning opportunity”. *Educational Data Mining 2010*. Citeseer. 2010.

[Leó16] León, M. et al. “Visualising the MOOC experience: a dynamic MOOC dashboard built through institutional collaboration”. *Research Track* (2016), p. 461.

[Lew19] Lewis, D. G. R. et al. “Opportunities for educational innovations in authentic project-based learning: understanding instructor perceived challenges to design for adoption”. *Educational Technology Research and Development* **67**.4 (2019), pp. 953–982.

[Li17] Li, X. et al. “Exploring n-gram features in clickstream data for MOOC learning achievement prediction”. *International Conference on Database Systems for Advanced Applications*. Springer. 2017, pp. 328–339.

[Lin09] Lin, F.-R. et al. “Discovering genres of online discussion threads via text mining”. *Computers & Education* **52**.2 (2009), pp. 481–495.

[Liu16] Liu*, Z. et al. “MOOC Learner Behaviors by Country and Culture; an Exploratory Analysis”. *Proceedings of the 2016 Conference on Educational Data Mining*. Ed. by Barnes, T. et al. International Educational Data Mining Society, 2016.

[Mac] Machov, K. & Rakuinec, T. “Social web mapping dynamic analysis of social networks”. *SOCIAL MEDIA* **6** ().

[Mak10] Mak, S. et al. “Blogs and forums as communication and learning tools in a MOOC”. *Proceedings of the 7th International Conference on Networked Learning 2010*. University of Lancaster. 2010, pp. 275–285.

[Mal11] Maldonado, R. M. et al. “Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop”. *Educational Data Mining 2011*. 2011.

[Mar16] Marzouk, Z. et al. “Generating Learning Analytics to Improve Learners’ Metacognitive Skills Using nStudy Trace Data and the ICAP Framework.” *LAL@ LAK*. 2016, pp. 11–16.

[Mat19] Matthews, P. “Human-In-The-Loop Topic Modelling: Assessing topic labelling and genre-topic relations with a movie plot summary corpus”. *The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization*. Ergon-Verlag. 2019, pp. 181–207.

[Maz03] Mazza, R. & Dimitrova, V. “CourseVis: Externalising student information to facilitate instructors in distance learning”. *Proceedings of the International conference in Artificial Intelligence in Education*. Sydney, Australia. 2003, pp. 117–129.

[Maz07] Mazza, R. & Dimitrova, V. “CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses”. *International Journal of Human-Computer Studies* **65**.2 (2007), pp. 125–139.

[McA10] McAuley, A. et al. “The MOOC model for digital practice” (2010).

[McL11] McLaren, B. M. et al. “A politeness effect in learning with web-based intelligent tutors”. *International Journal of Human-Computer Studies* **69**.1-2 (2011), pp. 70–79.

[Mik13] Mikolov, T. et al. “Distributed representations of words and phrases and their compositionality”. *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[Mor11] Moran, M. et al. “Teaching, Learning, and Sharing: How Today’s Higher Education Faculty Use Social Media.” *Babson Survey Research Group* (2011).

[Mvu12] Mvududu, N. & Thiel-Burgess, J. “Constructivism in practice: The case for English language learners”. *International Journal of Education* **4**.3 (2012), pp. 108–118.

[New90] Newman, R. S. "Children's help-seeking in the classroom: The role of motivational factors and attitudes." *Journal of educational psychology* **82**.1 (1990), p. 71.

[New94] Newman, R. S. "Adaptive help seeking: A strategy of self-regulated learning." (1994).

[NA19] Ngoc Anh, B. et al. "A computer-vision based application for student behavior monitoring in classroom". *Applied Sciences* **9**.22 (2019), p. 4729.

[Pak11] Pak, J. & Zhou, L. "A social network based analysis of deceptive communication in online chat". *Workshop on E-Business*. Springer. 2011, pp. 55–65.

[Pal20] Palese, B. & Piccoli, G. "Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-standard Sets". *Communications of the Association for Information Systems* **47**.1 (2020), p. 16.

[Pal08] Palmer, S. et al. "Does the discussion help? The impact of a formally assessed online discussion on final student results". *British Journal of Educational Technology* **39**.5 (2008), pp. 847–858.

[Pan02] Pang, B. et al. "Thumbs up?: sentiment classification using machine learning techniques". *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics. 2002, pp. 79–86.

[Pen14] Peng, L. et al. "A case study of college teacher's politeness strategy in EFL classroom". *Theory and practice in language studies* **4**.1 (2014), p. 110.

[Pen20] Peng, X. et al. "Topic tracking model for analyzing student-generated posts in SPOC discussion forums". *International Journal of Educational Technology in Higher Education* **17**.1 (2020), pp. 1–22.

[Pia77] Piaget, J. *The development of thought: Equilibration of cognitive structures.(Trans A. Rosin)*. Viking, 1977.

[Pra16] Prabowo, D. A. et al. "TF-IDF-Enhanced Genetic Algorithm Untuk Extractive Automatic Text Summarization". *Jurnal Teknologi Informasi dan Ilmu Komputer* **3**.3 (2016), pp. 208–215.

[Qu15] Qu, H. & Chen, Q. "Visual analytics for MOOC data". *IEEE computer graphics and applications* **35**.6 (2015), pp. 69–75.

[Rab14] Rabbany, R. et al. "Collaborative learning of students in online discussion forums: A social network analysis perspective". *Educational data mining*. Springer, 2014, pp. 441–466.

[Ram14a] Ramesh, A. et al. "Understanding MOOC discussion forums using seeded LDA". *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*. 2014, pp. 28–33.

[Ram14b] Ramesh, A. et al. “Understanding MOOC Discussion Forums using Seeded LDA”. 2014, pp. 28–33.

[Ram03] Ramos, J. et al. “Using tf-idf to determine word relevance in document queries”. *Proceedings of the first instructional conference on machine learning*. Vol. 242. 2003, pp. 133–142.

[Ree10] Reed, M. S. et al. “What is social learning?” *Ecology and society* **15**.4 (2010).

[Reu01] Reusser, K. “Co-constructivism in educational theory and practice” (2001).

[Reu15] Reusser, K. & Pauli, C. “Co-constructivism in educational theory and practice”. *International encyclopedia of the social & behavioral sciences* **3** (2015), pp. 913–917.

[Rie14] Rienties, B. et al. “To let students self-select or not: that is the question for teachers of culturally diverse groups”. *Journal of Studies in International Education* **18**.1 (2014), pp. 64–83.

[Rob15] Robinson, A. “Exploring class discussions from a massive open online course (MOOC) on cartography”. English (US). *Modern Trends in Cartography - Selected Papers of CARTOCON 2014*. Ed. by Brus, J. et al. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, 2015, pp. 173–182.

[Rog10] Rogers, E. M. *Diffusion of innovations*. Simon and Schuster, 2010.

[Ros14] Rosé, C. P. et al. “Social factors that contribute to attrition in MOOCs”. *Proceedings of the first ACM conference on Learning@ scale conference*. ACM. 2014, pp. 197–198.

[Ros09] Rosen, D. & Corbit, M. “Social network analysis in virtual environments”. *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. 2009, pp. 317–322.

[Ros11] Rosen, D. et al. “Social and semantic network analysis of chat logs”. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. 2011, pp. 134–139.

[RV15] Ruipérez-Valiente, J. A. et al. “ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform”. *Computers in Human Behavior* **47** (2015), pp. 139–148.

[Rya97] Ryan, A. M. & Pintrich, P. R. “” Should I ask for help?” The role of motivation and attitudes in adolescents’ help seeking in math class.” *Journal of educational psychology* **89**.2 (1997), p. 329.

[Sar20] Sari, A. R. et al. “MOOC instructor designs and challenges: what can be learned from existing MOOCs in Indonesia and Malaysia?” *Asia Pacific Education Review* **21**.1 (2020), pp. 143–166.

[Sch16] Schuck, S. et al. “Designing an iPad app to monitor and improve classroom behavior for children with ADHD: iSelfControl feasibility and pilot studies”. *PloS one* **11**.10 (2016), e0164229.

[Sea14] Seaton, D. T. et al. “Who Does What in a Massive Open Online Course?” *Commun. ACM* **57**.4 (2014), pp. 58–65.

[Sed14] Sedgwick, P. “Spearman’s rank correlation coefficient”. *BMJ: British Medical Journal (Online)* **349** (2014).

[Sha15] Sharma, K. et al. “Identifying Styles and Paths toward Success in MOOCs.” *International Educational Data Mining Society* (2015).

[She11] Shepherd, M. A. “Effects of ethnicity and gender on teachers’ evaluation of students’ spoken responses”. *Urban Education* **46**.5 (2011), pp. 1011–1028.

[Shi15] Shi, C. et al. “VisMOOC: Visualizing video clickstream data from massive open online courses”. *2015 IEEE Pacific visualization symposium (PacificVis)*. IEEE. 2015, pp. 159–166.

[SB17] Shirvani Boroujeni, M. et al. “Dynamics of MOOC Discussion Forums”. *7th International Learning Analytics and Knowledge Conference (LAK17)*. EPFL-CONF-223718. 2017.

[Shu12] Shum, S. B. & Ferguson, R. “Social learning analytics”. *Journal of educational technology & society* **15**.3 (2012), pp. 3–26.

[Sie12] Siemens, G. & Baker, R. S. d. “Learning analytics and educational data mining: towards communication and collaboration”. *Proceedings of the 2nd international conference on learning analytics and knowledge*. 2012, pp. 252–254.

[Sin14a] Sinha, T. et al. “Capturing” attrition intensifying” structural traits from didactic interaction sequences of MOOC learners”. *arXiv preprint arXiv:1409.5887* (2014).

[Sin14b] Sinha, T. et al. “Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions”. *arXiv preprint arXiv:1407.7131* (2014).

[Smi17] Smith, A. J. et al. “My digital hand: A tool for scaling up one-to-one peer teaching in support of computer science learning”. *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. 2017, pp. 549–554.

[Son17] Song, J. et al. “TOLA: Topic-oriented learning assistance based on cyber-physical system and big data”. *Future Generation Comp. Syst.* **75** (2017), pp. 200–205.

[Spe14] Speck, J. et al. “ForumDash: Analyzing online discussion forums”. *Proceedings of the first ACM conference on Learning@ scale conference*. 2014, pp. 139–140.

[Sta06] Stahl, G. et al. “Computersupported collaborative learning: An historical perspective, 2006”. *Cambridge handbook of the learning sciences* (2006), pp. 409–426.

[Sta00] Stamatatos, E. et al. "Automatic text categorization in terms of genre and author". *Computational linguistics* **26**.4 (2000), pp. 471–495.

[Ste17] Stepanek, N. R. & Dorn, B. "Automated Analysis of Lecture Video Engagement Using Student Posts". *International Conference on Artificial Intelligence in Education*. Springer. 2017, pp. 565–569.

[SM14] Stephens-Martinez, K. et al. "Monitoring moocs: which information sources do instructors value?" *Proceedings of the first ACM conference on Learning@ scale conference*. 2014, pp. 79–88.

[Ste11] Stewart, S. A. & Abidi, S. S. R. "Using social network analysis to study the knowledge sharing patterns of health professionals using web 2.0 tools". *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer. 2011, pp. 335–352.

[Ste07] Steyvers, M. & Griffiths, T. "Probabilistic topic models". *Handbook of latent semantic analysis* **427**.7 (2007), pp. 424–440.

[Tat08] Tatar, D. et al. "Scaling up innovative technology-based mathematics". *The Journal of the Learning Sciences* **17**.2 (2008), pp. 248–286.

[Tay14] Taylor, C. et al. "Likely to stop? predicting stopout in massive open online courses". *arXiv preprint arXiv:1408.3382* (2014).

[Tep11] Teplovs, C. et al. "Generating predictive models of learner community dynamics". *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. 2011, pp. 147–152.

[The10] Thelwall, M. et al. "Sentiment strength detection in short informal text". *Journal of the American Society for Information Science and Technology* **61**.12 (2010), pp. 2544–2558.

[Tuc14] Tucker, C. et al. "Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes". *Computers in Education Journal* **5**.4 (2014), pp. 84–95.

[VD01] Van Dijk, L. et al. "Interactive lectures in engineering education". *European Journal of Engineering Education* **26**.1 (2001), pp. 15–28.

[Vie05] Viera, A. J., Garrett, J. M., et al. "Understanding interobserver agreement: the kappa statistic". *Fam Med* **37**.5 (2005), pp. 360–363.

[Vig17] Vigentini, L. et al. "Overcoming the MOOC data deluge with learning analytic dashboards". *Learning analytics: Fundaments, applications, and trends*. Springer, 2017, pp. 171–198.

[Vu15] Vu, D. et al. "Relational event models for social learning in MOOCs". *Social Networks* **43** (2015), pp. 121–135.

[Vyg78] Vygotsky, L. S. & Cole, M. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

[Wag01] Wagstaff, K. et al. “Constrained k-means clustering with background knowledge”. *Icml*. Vol. 1. 2001, pp. 577–584.

[Wan08] Wan, X. & Yang, J. “Multi-document summarization using cluster-based link analysis”. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, pp. 299–306.

[Wan12] Wang, X. et al. “Text clustering based on the improved TFIDF by the iterative algorithm”. *2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM)*. IEEE. 2012, pp. 140–143.

[Wan15a] Wang, X. et al. “Investigating How Student’s Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains.” *International Educational Data Mining Society* (2015).

[Wan16a] Wang, X. et al. “Towards triggering higher-order thinking behaviors in MOOCs”. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM. 2016, pp. 398–407.

[Wan16b] Wang, X. et al. “Towards Triggering Higher-order Thinking Behaviors in MOOCs”. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. LAK ’16. Edinburgh, United Kingdom: ACM, 2016, pp. 398–407.

[Wan15b] Wang, Y. & Baker, R. “Content or platform: Why do students complete MOOCs?” *Journal of Online Learning and Teaching* **11**.1 (2015), p. 17.

[Wen14] Wen, M. et al. “Sentiment Analysis in MOOC Discussion Forums: What does it tell us?” *Educational data mining 2014*. Citeseer. 2014.

[Whi15] Whitehill, J. et al. “Beyond prediction: First steps toward automatic intervention in MOOC student stopout”. *Available at SSRN 2611750* (2015).

[Wis18] Wise, A. F. & Cui, Y. “Unpacking the relationship between discussion forum participation and learning in MOOCs: content is key”. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM. 2018, pp. 330–339.

[Wis16] Wise, A. F. et al. “Bringing order to chaos in MOOC discussion forums with content-related thread identification”. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM. 2016, pp. 188–197.

[Xha16] Xhakaj, F. et al. “How teachers use data to help students learn: Contextual inquiry for the design of a dashboard”. *European Conference on Technology Enhanced Learning*. Springer. 2016, pp. 340–354.

[Xu19] Xu, Y. et al. “What do You Want? Applying Deep Learning Models to Detect Question Topics in MOOC Forum Posts”. *In Proceedings of the 2019 KDD Work-*

shop on Deep Learning for Education (DL4Ed), held in conjunction with 2019 KDD Anchorage, Alaska, U.S.A. Ed. by Lan, A. S. et al. 2019.

[Yan14] Yang, D. et al. "Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs". *The International Review of Research in Open and Distributed Learning* **15**.5 (2014).

[Ye14] Ye, C. & Biswas, G. "Early Prediction of Student Dropout and Performance in MOOCSs Using Higher Granularity Temporal Information." *Journal of Learning Analytics* **1**.3 (2014), pp. 169–172.

[Yu19] Yu, C.-H. et al. "Predicting learning outcomes with MOOC clickstreams". *Education sciences* **9**.2 (2019), p. 104.

[Zha02] Zha, H. "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering". *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM. 2002, pp. 113–120.

[Zha07] Zhang, J. et al. "Expertise networks in online communities: structure and algorithms". *Proceedings of the 16th international conference on World Wide Web.* ACM. 2007, pp. 221–230.

[Zhe16] Zheng, S. et al. "Ask the instructors: Motivations and challenges of teaching massive open online courses". *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* 2016, pp. 206–221.

[Zhu16] Zhu, M. et al. "Longitudinal engagement, performance, and social connectivity: a MOOC case study using exponential random graph models". *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.* ACM. 2016, pp. 223–230.

[Zim94] Zimmerman, B. J. "Dimensions of academic self-regulation: A conceptual framework for education". *Self-regulation of learning and performance: Issues and educational applications* **1** (1994), pp. 33–21.