Spring 2023

Learning Analytics Through Machine Learning and Natural Language Processing

Bokai Yang

LEARNING ANALYTICS THROUGH MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

by

Bokai Yang

Master of Science
University of Florida, 2018

Bachelor of Engineering
University of Electronic Science and Technology of China, 2016

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Computer Science

College of Engineering & Computing

University of South Carolina

2023

Accepted by:

John R. Rose, Major Professor

Marco Valtorta, Committee Member

Csilla Farkas, Committee Member

Jijun Tang, Committee Member

Hengtao Tang, Committee Member

Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School

ABSTRACT

The increase of computing power and the ability to log students' data with the help of the computer-assisted learning systems has led to an increased interest in developing and applying computer science techniques for analyzing learning data. To understand and investigate how learning-generated data can be used to improve student success, data mining techniques have been applied to several educational tasks. This dissertation investigates three important tasks in various domains of educational data mining: learners' behavior analysis, essay structure analysis and feedback providing, and learners' dropout prediction. The first project applied latent semantic analysis and machine learning approaches to investigate how MOOC learners' longitudinal trajectory of meaningful forum participation facilitated learner performance. The findings have implications on refining the courses' facilitation methods and forum design, helping improve learners' performance, and assessing learners' academic performance in MOOCs. The second project aims to analyze the organizational structures used in previous ACT test essays and provide an argumentative structure feedback tool driven by deep learning language models to better support the current automatic essay scoring systems and classroom settings. The third project applied MOOC learners' forum participation states to predict dropouts with the help of hidden Markov models and other machine learning techniques. The results of this project show that forum behavior can be applied to predict dropout and evaluate the learners' status. Overall, the results of this

dissertation expand current research and shed light on how computer science techniques could further improve students’ learning experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Since data mining was first applied to improve learning environments in 1995, the interest in educational data mining and educational systems keeps growing and makes educational data mining a growing research community (Romero & Ventura, 2007). The increase in computing power and the ability to log students’ data with the help of the new learning systems has led to an increased interest in developing and applying computer science techniques for analyzing the learning data. Novel computer-assisted learning platforms generate large amounts of learning data and it is essential to put these data in context, understand the behavior behind the data, and use the data to reflect and predict learners’ behavior to better enhance their learning. To understand and investigate how learning generated data can contribute back to learners’ learning, data mining and artificial intelligence methods have been applied to several education tasks in different learning platforms (e.g., college e-learning systems, MOOC and so on). The tasks include behavior analysis, behavior prediction, sentiment analysis, learning recommendation systems, automated essay scoring and so on. In this dissertation, we studied three important tasks in domains of educational data mining: learners’ behavior analysis, essay structure analysis and feedback providing, and learner dropout prediction. We introduce these problems below respectively.

The first problem we investigated is to apply traditional machine learning and natural language processing techniques to identify the MOOC learners' meaningful participation patterns. An effective discussion forum is essential for both MOOC facilitators and learners. Identifying the meaningful contents can help MOOC facilitators fast locate the course/system related questions and increase the learners' learning experience. However, with the massive enrollment, the MOOC forums are filled with chaos. Course-related messages are flooded by course-unrelated information which makes it hard for facilitators and learners to find the information they need. Thus, to address this problem, we first applied latent semantic analysis and a decision tree model to classify the forum posts into topic-related posts and topic-unrelated posts. Then with the posts categorized, the learners' forum participation patterns were extracted. The inferential statistical results showed that learners' performance (i.e., grade and dropout) was associated with their longitudinal pattern of meaningful forum participation. The first project is finished and was published by Computers & Education.

For the second problem we analyzed the organizational structures used in previous ACT test essays with the help of deep learning massive language models. In addition, we developed an argumentative structure feedback tool to better support the current automatic essay scoring systems and classroom settings. Writing argumentative essays is a critical component of students' learning. As the ratio of teachers to students declined, the manual evaluation process becomes more and more time and effort consuming. In recent years, some automatic essay scoring tools have emerged, aiming to save manual grading effort and grade students' essay without human interference. However, most of the current AES tools provide a holistic score of the essay which summarizes the quality of an essay as a whole which is far not enough to be applied in classroom settings. Thus, it is essential to provide feedback for both the teachers and students. To fill this gap, we analyzed the association between organizational structures and scores in previous tests essays. With the analyzed results, we designed and developed a feedback tool which aims to help students improve their argument writing skills with a better understanding of argumentative elements and structures.

The third problem was to explore the association between forum participation and MOOC learners' dropout and how learners' forum participation states could be applied to increase the performance of the current dropout prediction models. Most of the previous research applied forum participation as a feature to predict dropout through a single dimension, which is insufficient. In this study, with the help of Hidden Markov Model (HMM), we analyzed MOOC forum behavior and the association between dropout through multiple dimensions, including a quantitative dimension, a content dimension, and a temporal dimension. In order to explore the power of learners' forum participation as a feature in predicting dropouts, we further developed several dropout prediction models with widely applied clickstream features and tested if the performance of the models could be improved with the learners' forum participation HMM states. The results show that forum participation is an important feature for predicting dropout. Integrating continuous forum participation states as features can significantly increase the accuracy of the dropout prediction models.

1.2 Literature Review

Formal statistical inference is assumption driven. It first forms a hypothesis and then tests against the data. In contrast, Data mining is discovery driven and the hypothesis is automatically extracted from the data (Romero & Ventura, 2007). As shown in Figure 1.1, Data mining comprises both statistics and artificial intelligence (i.e., machine learning and deep learning). Educational data mining comprises education, statistics, and artificial intelligence. This researcg aims to explore how educational data mining could be applied and enhance students learning through multiple projects.

Educational Data Mining (EDM) is a research area focused on developing methods to explore the data that come from educational environments (Bakhshinategh et al., 2018). To be specific, EDM applies data mining techniques to the educational environment produced datasets and tends to answer important educational questions through the analysis results (Romero & Ventura, 2013). The tasks of EDM include predicting student performance, detecting and understanding students' behaviors, providing reports to facilitators and help to learners, group and profile learners and so on.

The starting point of applying educational data mining is the evaluations of students' usage of an e-learning system (Tsantis & Castellani, 2001). After that, traditional machine learning methods have been proposed to explore and understand the data generated by the e-learning courses (e.g., Talavera & Gaudioso, 2004). For example, in early studies, Tang et al. (2000) uses a clustering-based method to promote group-based collaborative learning and to provide incremental learner diagnosis for web learning. Minaei-Bidgoli and Punch (2003) extracted features from logged data and predict students' final grades. Talavera and Gaudioso (2004) propose to mine student data through clustering to discover students’ behavior patterns. Until recent years, machine learning approaches are still widely applied since it can not only finish the prediction tasks but also be used to analyze and explain the data (e.g., Jeong et al., 2010; Moreno-Marcos et al., 2020; Wang et al., 2015).

With the development of machine learning, deep learning techniques have been applied into educational data mining. Most of the prediction EDM tasks are covered by DL approaches. For example, predicting students’ performance (e.g., Lin & Chi, 2017; Okubo et al., 2017; Tang et al., 2016) and predicting students’ dropout (e.g., Fei & Yeung, 2015; Wang et al., 2017; Xing & Du, 2018). However, deep learning approaches require large amounts of information to train the model and there are only a few large datasets that have been developed and available (Hernández-Blanco et al., 2019). In addition, works have been argued that traditional machine learning techniques could also achieve similar performance under fairer scenario (e.g., Lalwani & Agrawal, 2017; Hernández-Blanco et al., 2019; Khajah et al., 2016; Mao et al., 2018; Wilson et al., 2016; Xiong et al., 2016; Yeung & Yeung, 2018).

Automatic essay scoring is a specific research area within educational data mining. AES aims to grade students’ essays with the help of machine learning and natural-language processing. Similar to EDM, AES involves from early traditional machine learning (e.g., Dascalu et al., 2017; Persing & Ng, 2015; Rahimi et al., 2017; Zhang et al., 2019) approaches to deep learning approaches (Dong et al., 2017; Taghipour & Ng, 2016; Uto et al., 2020; Alikaniotis et al., 2016). Early studies applied feature engineering and considered multiple features that could influence the essay score. For example, length-based features (Yannakoudakis & Briscoe, 2012), lexical features (Phandi et al., 2015; Zesch et al., 2015), Prompt-relevant features (Klebanov et al., 2016), Argumentation features (Ghosh et al., 2016; Persing and Ng, 2015; Wachsmuth et al., 2016) and so on. Other than machine learning approaches, researchers tend to apply deep learning techniques to evaluate the essay without feature engineering (e.g., Alikaniotis et al., 2016; Dong et al., 2017; Taghipour & Ng, 2016). Neural networks have the advantages of modeling complex patterns in data and can be applied to predict the score without manual engineering features (Taghipour & Ng, 2016). However, to provide more detailed feedback for educational and classroom setting purposes, deep learning approaches also include different features into consideration. For example, Mim et al. (2019) proposed a DNN model to predict both coherence and argument strength scores of an essay. Hussein et al. (2020) proposed a model that not only provides a holistic score but also four traits' scores to the input essay. Their study aims to provide adaptive feedback to learners according to the traits scores.

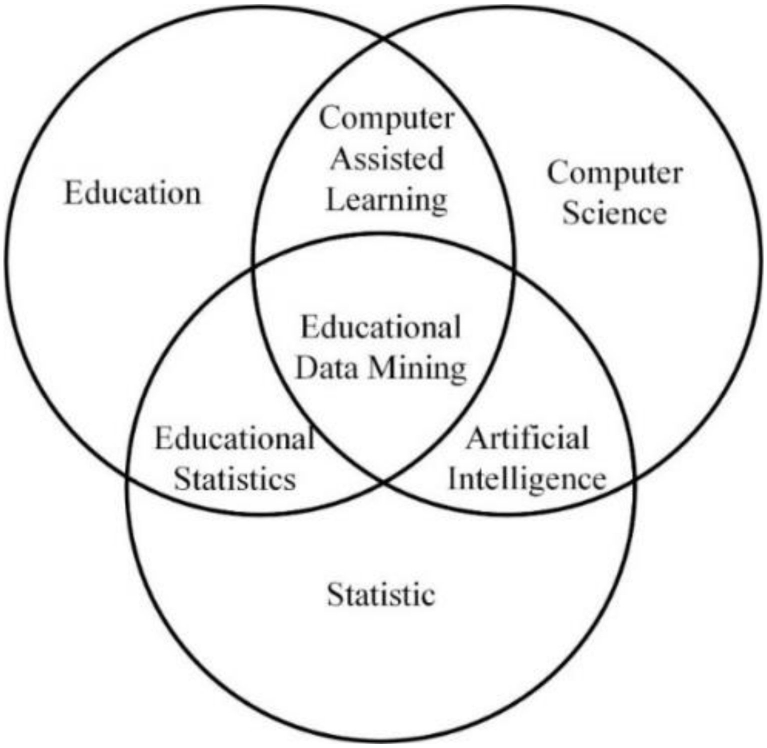


Figure 1.1 Areas related to educational data mining

CHAPTER 2

TEMPORAL ANALYSIS OF MOOC LEARNERS' FORUM PARTICIPATION

2.1 Introduction

An effective discussion forum is critical to the effectiveness of Massive Open Online Courses (MOOCs) for both instructors and learners. Interaction in forums has been the primary avenue for MOOC instructors to monitor how learning proceeds in the course (Jiang et al., 2015). On the other hand, learners participate in forum interactions to seek necessary support and feedback especially given that the dramatic instructor-student ratio in MOOCs (Stephens-Martinez et al., 2014). However, the massive number of learners contributes to an upsurge of posts in MOOC forums, a large portion of which are irrelevant to course topics (Brinton et al., 2014). This unexpected chaos results in information overload for learners, which makes it challenging to locate relevant posts (Wise et al., 2017). Thus, resolving the challenge of information overload in forums is vital for the effectiveness of MOOCs.

To untangle the chaos in MOOC forums, the primary step is to identify posts that are relevant to the course topic (Wise et al., 2017). Recently, the application of learning analytic techniques has enriched options for classifying forum posts (e.g., Agrawal et al., 2015; Brinton et al., 2014; Jiang et al., 2015; Wise et al., 2017). For instance, Wise et al. (2017) proposed a model using linguistic features to distinguish relevant from non-relevant posts. Classifying forum posts by their relevance to the course contents enables learners and instructors to promptly access relevant posts in MOOCs and opens the possibility for fruitful interactions and expanded insights beyond their knowledge (Wise et al., 2017). In addition, identifying topic-relevant posts in the forums and their impact on learner performance can provide insightful implications on supporting learners in MOOCs.

It is noteworthy that learner activities in discussion forums vary remarkably over time (Molenaar, 2014). Learners' forum activities (e.g., posting and commenting) are closely related to learner performance (e.g., grades and retentions), but posting and commenting activities at different time periods during the semester may have a different degree of influence on their performance (Tang et al., 2018; Xing et al., 2019). For example, Tang et al. (2018) found that learners' consistent forum activities lead to high course grade in MOOCs, but the influence of gradually disengaging forum participation on course grades is limited. Understanding the association between topic-related forum posts and learner performance in MOOCs thus required consideration of the temporal dimension of learners' content-related forum participation. Prior studies tapped into the chaos in discussion forums using machine-oriented methods to quantify the total number of meaningful posts and irrelevant posts in online forums (Wise et al., 2017; Xing et al., 2019). These methods seldom consider the variation of learners' posting and commenting activities. A lack of a temporal account of learners' meaningful posts and comments results in a gap of understanding about whether posting and/or commenting with topic-related content facilitate learner performance.

The purpose of this research is threefold. First, this research project seeks to use machine learning methods to classify relevant and irrelevant posts. Second, this research project describes the longitudinal trajectory of online learners' topic-related forum participation in MOOCs. Third, this research project is intended to investigate how learners' longitudinal trajectory of topic-related forum participation is associated with their course performance. We sought to answer the following three research questions:

RQ1: What are the longitudinal patterns of learners' meaningful forum participation over time?

RQ2: How do learners' course grades differ by their longitudinal patterns of meaningful forum participation?

RQ3: How does learners' longitudinal pattern of meaningful forum participation associated with their course retention?

2.2 Related Works

2.2.1 Chaos in MOOC Discussion Forums

Discussion forums and associated activities have become increasingly central to MOOC instruction as the primary setting for collaborative learning (Kellogg & Edelmann, 2015). Collaborative learning is an integral component of the learner's experience in MOOCs. MOOCs are the latest incarnation of online courses that bring together massive number of learners with significant variance as such their unique life experiences and sociocultural beliefs (Gillani & Eynon, 2014). Enrolled learners connect with one another and exchange ideas and/or knowledge so that collaborative learning results in an increase of both individual and collective knowledge in MOOCs. On the other hand, collaborative learning provides learners an alternative option to offset the lack of instructional support in MOOCs. An equivalent level of instructional feedback and supports as that in the traditional online courses in a college is deemed unlikely in

MOOCs due to a much larger enrollment (Kellogg et al., 2014). Without instructional support, learners might wrestle with such challenges as difficulty in understanding the course content, and the challenges might even lead to learners' dropping out.

Chaos has been widely investigated in MOOCs with the hope of identifying course-relevant posts and providing learners with prompt access to desired information. Wise et al. (2017) have summarized methods used to detect content-relevant posts in three various perspectives, such as machine-oriented (e.g., Agrawal et al., 2015), instructor-oriented (e.g., Jiang et al., 2015), and learner-oriented methods (e.g., Brinton et al., 2014). Machine-oriented methods mainly use automatic tools to classify forum posts and provide predesigned support based on the analysis of prior datasets of learner traces (Agrawal et al., 2015). Instructor-oriented methods mainly focus on analysis based on instructor intervention history and small talk threads of prominent learners (e.g., Jiang et al., 2015). The Learner-centered approach mainly depends on the use of various tools to intervene in learner interaction with forums, such as thread recommendation tools (Yang et al., 2014) or topic modelling. However, it is worth noting that all three methods do not consider temporal variation in meaningful learner engagement and potentially shape the relationship between these two variables. Therefore, tapping into the temporal perspective of meaningful online forum engagement is critical. Building upon existing outcomes, this research mainly focused on connecting temporal analysis and machine learning to accurately investigate the relationship discussed above.

2.2.2 Educational Data Mining in MOOC Discussion Forums

At the forum content level, there is a wealth of research focused on clustering and classifying the forum data in MOOCs through different data mining methods and with different foci. For instance, Speck et al. (2014) built a plugin called ForumDash which applied LSA to reveal discussion topic clusters. After applying ForumDash on a physics MOOC, three discussion topics are identified, i.e., homework problems and answers, questions about specific syllabus units, and other physics MOOCs offered. In a different study, Brinton et al. (2014) analyzed the forum activity, extracted topics, and categorized discussion threads into “small-talk”, “course logistics”, and “course specific” categories through naïve Bayes and support vector machine classifiers. Their results showed that the participation of the teaching staff could increase the discussion volume, but it did not reduce the rate of decline in participation.

In addition to addressing topics and threads, educational data science researchers also focused on investigating posts and comments in MOOC forums. Ezen-Can et al. (2015) proposed a k-medoids clustering approach and clustered the forum posts into seven clusters. All these studies focused on extracting topics, but they did not examine individual learner behavior. Through a two-stage content analysis approach, Wang et al. (2015) classified learners’ posts as on-task and off-task and assigned the posts into three categories (Active, Constructive, and Interactive) using logistic regression. Their results identified the relationships between learners’ discussion behavior and their learning. However, they did not analyze discussion behavior over time; and they did not consider the behavior of the learners who had dropped the course.

2.3 Methodology

2.3.1 Data Sources

In this work, our datasets included anonymized learners' grades and discussion forum traces from a MOOC *Introduction to Art: Concepts & Techniques*. This course lasted for seven weeks with up to 69,867 learners enrolled. The dataset recorded enrolled learners' demographic information, grades, clickstream data, peer reviews, and forum logs. The forum contained threads, posts, and comments. More specifically, a thread includes a group of posts which allows learners to communicate on a new topic; a post is a message responding to a thread; and a comment is a reply to a post. In total, there are 29,876 posts and 22,292 comments in the forum data. We did not distinguish between posts and comments in this research to determine the relevance of the content to the topic. Thus, the word "posts" in the following sections indicates both posts and comments. Figure 2.1 gives the number of posts over the seven weeks. The number of active learners is shown in Figure 2.2.

2.3.2 Topic-related and Topic-unrelated Posts

Learners' meaningful forum participation is embodied through posting content-relevant posts to the discussion forums. Specifically, we labeled forum posts relevant to course topics (e.g., self-introductions, discussions about artwork, questions and comments about the course contents, and assignment questions) as topic-related posts. Course-irrelevant topics included technical problems, course administration issues, course evaluations, complaints and so on. The posts with content in course-irrelevant topics were labeled topic-unrelated posts. Examples of topic-related and topic-unrelated posts are shown in Table 2.1.

The original dataset was generated by two researchers with experience in educational research and machine learning expertise. After removing the posts with only symbols (e.g., emoji and punctuations) and invalid posts (e.g., non-English posts, missing links), 4,108 posts were chosen randomly and independently coded manually by these two researchers. The Kappa value (k=0.82) was calculated, confirmed that the two researchers reached an acceptable level of interrater reliability on the preliminary codes. The two researchers then discussed the disparities in codes and reached an agreement on all the 4,108 posts. Finally, a total of 3,526 posts were labeled as topic-related and the other 582 posts were topic-unrelated.

2.3.3 Dataset

The original dataset was then partitioned into a training set and a test set to be used to train and test the classification model. Table 2.2 shows the dataset we used to conduct our study. The training set was used to construct the LSA model and train the decision tree model. The training dataset is comprised of 1232 labeled posts which contains 754 topic-related posts and 432 topic-unrelated posts. We note that 46 posts in the training set were removed by the LSA text preprocess procedure. The test set for the decision tree model contains the remaining topic-unrelated posts from the full dataset and some randomly selected topic-related posts that could pass the text preprocessing procedure. The test set contains 443 posts, with 293 topic-related and 150 topic-unrelated posts.

2.3.4 Data Processing and Analysis Procedures

A three-staged sequential analysis was conducted, including latent semantic analysis, machine learning analysis via the decision tree model, and longitudinal K-means cluster analysis. Figure 2.3 presents the general procedures of our system.

The first stage creates the LSA model and computes the vector coordinates of each post in the reduced 2-dimensional space. To preprocess the posts, we applied the following series of text-mining procedures: (1) tokenized posts with non-letter separators; (2) removed stop words such as “a”, “the”, “is”, “are”, and so on; (3) applied lemmatization to remove word variations, e.g., “drew” and “draws” were transformed to “draw”; (4) performed part of speech (POS) analysis to remove all the words other than open-class words (i.e. nouns, verbs, adjectives and adverbs) through tagging; (5) applied word reduction and duplication based on the word’s occurrence and an topic-unrelated dictionary. One issue with using LSA is the need to tune its semantic resolution. This is done by finding the appropriate number of dimensions. A small number of dimensions may cause a poor performance and a large number of dimensions may make the performance the same as word matching (Dumais 2004). To reduce the number of dimensions and improve the sensitivity of the LSA model, we removed the words that occurred fewer than twenty times in the LSA set. However, since the proportion of irrelevant posts in the original dataset was low, the topic-unrelated features (i.e., some frequent words in topic-unrelated posts) may be reduced during the process of dimensionality reduction. To keep a proper scale of topic-related and topic-unrelated terms in the LSA occurrence matrix, we manually set the ratio of topic-related posts and topic-unrelated posts, and further duplicated some topic-unrelated features (59 features in total) for all the input posts. The first four procedures of the LSA model depicted in Figure 2.3 were achieved with the help of OpenNLP (OpenNLP, 2011). After preprocessing the LSA set, 46 posts (3.7%) were removed for lack of length. 433 features were extracted and used to construct an LSA model using quanteda (Benoit et. al., 2018). We then used this LSA model to compute the reduced 2-dimensional vector coordinates of each post.

At the second stage of our process, the reduced 2-dimensional results of the posts in LSA set were used to train a decision tree model using Weka (Eibe et. al., 2016). The decision tree model is used to classify the posts into two categories: topic-related and topic-unrelated.

Finally, after all posts were categorized, we analyzed post activity over time. The number of posts for each student in each week is calculated according to their relevance. We apply a longitudinal data cluster analysis using the K-means algorithm through the kml-package in R (Genolini & Falissard, 2010). To eliminate the outliners, we only considered the students who posted between 7 topic-related posts (at least one post per week) and 70 topic-related posts (at most 10 posts per week) over the seven-week period. The students' behavior patterns are then clustered into 5 clusters shown in the Results section.

To analyze the relationship between student meaningful forum participation and learner performance, we consider two performance metrics 1) course final grade and 2) learner retention. The final grade was the variable used to assess learner performance in major assessment activities such as quizzes and assignments. The grade was a continuous variable, ranging from 0 – 100, and students earning 70 or above were eligible to receive a course certificate. An analysis of variance (ANOVA) test was conducted on learners' longitudinal patterns of meaningful forum participation and their final grades to answer RQ2.

Learner retention was measured by course dropout. Several previous studies gave different definitions of dropout. Whitehill et al. (2017) proposed to use the students' grade as the measurement to decide learners' dropout. They labeled a student as a dropout if he/she did not accrue enough points to earn a certificate. Other than considering the grades of the learners, some research considered the timing of no learning activity as the dropout point (Chen et al., 2019; Zheng et al., 2020). Rather than considering single measurement, mixed measures have been proposed (Nagrecha et al. 2017; Moreno-Marcos et al. 2020). Nagrecha et al. (2017) used the definition that if a student has no interaction between some point during the course and the end of the course and he/she has not completed a certification, then the student is considered a dropout. Moreno-Marcos et al. (2020) set the inactive period to four weeks and considered submitting at least 80% of the assessments as the grade measurement.

Given the uniqueness of MOOCs, it is difficult to determine course dropouts through a single measurement. Some learners might be inactive in the last two weeks, but they still received enough points to earn a certificate. On the other hand, some learners may remain active until the last week of the course even though they did not receive enough points to earn a certificate. Thus, we adapted the dropout definitions of Nagrecha et al. (2017) and Moreno-Marcos et al. (2020) based on the configuration of our MOOC. We determined the course dropouts through two variables, last active time and final grade. We first determined the learners' activeness by judging whether the learners were still active in the last week. This was a binary variable where “0” denotes attrition from the course while “1” indicates students stayed engaged in the course until the final week of the course. With the definition of the learners’ activeness, we added one more variable which is final grade to decide the dropout status. If a learner received less than 80% of the final grade to get a certification (i.e., 56), and his/her activeness variable is 0, the learner is considered to be a dropout. A chi-square analysis was conducted to determine how each cluster of learners differed by their retention in the course.

2.3.5 Determining the Number of Clusters

The kml package provides several criteria to select the number of clusters. To make all the criteria comparable, the kml package computes the opposite of the criteria that should be minimized, which means the maximized results are preferred (Genolini & Falissard, 2015). After examining the plots of all the criteria, we decided to rely on the Ray-Turi criterion (Ray & Turi, 1999) and the Davies-Bouldin criterion (Davies & Bouldin, 1979), since both criteria minimize intra-cluster distance and maximum inter-cluster distance. As shown in Figure 2.4, both criteria agree on 5 clusters as the maximized results.

2.4. Results

2.4.1 Learners’ Meaningful Forum Participation

After feeding all the posts to the model, 38,239 posts (20,945 posts and 17,294 comments) were marked as topic-related. The other posts are either topic-unrelated, invalid (the posts were not posted in English or do not have meaningful content, e.g., post with an emoji) or the posts that cannot be analyzed (the posts do not contain the 433 features in the LSA model mentioned before). Figure 2.5 presents the results produced by our classification model with the numbers of students' topic-related posts and topic-unrelated posts over seven weeks. As can be observed, the number of topic-unrelated posts is roughly constant while the number of topic-related posts declines with an exponential curve.

Our models demonstrate reasonably good reliability in identifying related/unrelated posts. As shown in Table 2.3, the accuracy of our model is 0.76 with an AUC value of 0.83 and a Kappa value of 0.506. This is similar to the results of Wang et al. (2015).

2.4.2 Learners' Longitudinal Patterns

After eliminating outliers, 1326 students posted more than 7 topic-related posts in the seven-week course. As can be seen in Figure 2.6, the students were clustered into 5 clusters (A, B, C, D, E) according to their post activity trajectories. Table 2.4 shows the average number of posts per week for each cluster. The number of students in each cluster and their proportions are shown in Table 2.5. To further analyze the features of each cluster, we corelated students' behavior trajectories and their grades. Students' average grades and the standard deviations are also shown in Table 2.5.

2.4.3 Longitudinal Patterns and Learner Grades

Table 2.5 presents the descriptive statistics for each cluster's course grade. The ANOVA analysis results confirmed that there was a significant difference in course grades among the five clusters of learners, $F(4)=17.70$, $p<0.01$. Fisher's Least Significant Difference (LSD) analysis suggested the mean score of Cluster B was significantly lower than the other four clusters. In addition, Cluster D earned a significantly higher mean score than the other clusters except for Cluster E. The difference in mean scores between Cluster D and Cluster E was not significant. The mean score of Cluster C was significantly lower than that of Cluster E, but there was no significant difference in mean scores between Cluster A and Cluster C as well as between Cluster A and Cluster E.

2.4.4 Longitudinal Patterns and Learner Retention

The Chi-square analysis result ($\chi 2 = 65.71$; df = 4; $p < .01$) confirmed that learner retention in the course differed among the five clusters of learners. The effect size indicated that the difference was associated between the two variables, Cramer's V = 0.223, Phi = 0.223, $p < 0.01$. Specifically, the adjusted residuals for Clusters A, D, and E were larger than 2.0 (see Table 2.6), indicating that attrition from the course in those three clusters was significantly lower than would be expected if the null hypothesis were true, with a significance level of .05. In addition, the adjusted residuals for Clusters B, were smaller than -2.0 (see Table 2.6), indicating that attrition from the course in this cluster was significantly higher than would be expected if the null hypothesis were true, with a significance level of .05.

2.4.5 Cluster Characteristics

We identified five different types of learners (“light”, “transient”, “heavy starter”, “moderate and persistent”, and “entropic”) based on their posting frequency and related the learners’ posting behaviors to their grades.

Cluster A was labelled “light” learners. 43.4% of learners were in this cluster. This cluster includes the learners who posted lightly over the seven weeks. The small peek in week three was caused by the regularly posting leaners who dropped later. By observing the grades, we found that although these learners posted a low number of posts every week, their posting behavior still leads to a high possibility of getting a satisfactory grade.

However, from the variance analysis, it is obvious that the learners in cluster D and E are more likely to get a higher grade than the "light" learners in cluster A.

Cluster B was labelled "transient" learners. 36.6% of learners were in this cluster. These learners display remarkable posting activity in the first week. After the first week, around 30% of the learners in cluster B dropped the course, which resulted in a bimodal grade distribution. Those leaners who kept posting meaningful posts until the end of the course received good grades.

Cluster C was labelled "heavy starter" learners. This cluster had 7.47% of the participants. These learners posted a prominent number of posts in the first week and maintained a high degree of participation in the following two weeks. However, almost 30% of the learners in this cluster dropped the course in the first three weeks. In contrast to the learners in cluster B, the "heavy starter" posting frequency was a little bit higher over the seven weeks. Thus, most of the learners in this cluster either got a good grade or received 0.

Cluster D is labelled "moderate and persistent" learners. There were 6.49% learners in this cluster. We could easily conclude from Table 2.4 that the learners in cluster D stayed at a high level of participation during the entire course. They posted moderately every week. As can be observed from the variance analysis results, the leaners in cluster D earned a significantly higher mean score than the other clusters except for Cluster E. The leaners in cluster D also had the highest course retention rate. Only three students dropped the course after week 5 and almost all the other students received a high grade.

Cluster E is labelled “entropic” learners due to the pattern of initial higher energy state of posting gradually dissipating to a low energy state. In this study, 6.04% of the participants were in this cluster. Compared to the learners in cluster D, the learners in this cluster are more active in the first three weeks. However, their participation declined in the following weeks. Compared to the learners in cluster B and C, the learners in cluster E maintained longer meaningful forum participation (over first five weeks). The following weeks saw a significant decrease in their average forum participation. However, it did not lead to a significant loss of grade. Although the students in cluster E had a little bit lower mean score compared to the learners in cluster D, they still earned a significantly higher mean score than the learners in other clusters.

2.5. Discussion and Implications

The purpose of this study was to explore how learners’ meaningful forum participation related to their course performance. Using language processing and machine learning techniques, the research identified topic-related and topic-unrelated posts in MOOC forums. This was followed by temporal analysis which identified five clusters of learners by their longitudinal patterns of posting topic-related content in discussion forums. Statistical analysis indicated that learners’ longitudinal patterns of posting topic-related content were significantly associated with their course grades and course retention. The findings of this study provide empirical recommendations for MOOC course instructors and instructional designers to afford effective discussion forums.

Our study indicated the importance of learners’ meaningful forum participation in MOOCs. To gauge learners’ meaningful participation, we built a model to classify the topic-related and topic-unrelated posts in discussion forum. Previous studies have classified the posts through its relevance to the course content (Wang et al., 2015; Wise et al., 2017). In this study, the model we built is able to classify learners' meaningful participation in a more detailed way through their course topic relevance. That is to say, the classification model we proposed could be used to identify both the topic-related and topic-unrelated posts. Classifying forum posts by their relevance to the course content encourages learners' interaction and improves their learning gains (Wise et al., 2017). Thus, by applying this model to the MOOC setting, the instructor and the learners may filter course irrelevant information which reduce information overload, promote effective and efficient discussion forum interactions, and further prevent learners' attrition. In addition, based on our analysis, the ratio of the topic-unrelated posts to topic-related posts abruptly increased after week 3. This finding of this study suggests that course designers need to facilitate meaningful conversations to promote topic-related posts in the middle and the late phase of the MOOC.

In addition, our study supports the importance of identifying longitudinal patterns of learners' meaningful forum participation for predicting learner retention. Similar to the findings of Moreno-Marcos et al. (2020) who claimed that the best time to predict learners' dropout was in the second week of the course, we also noticed that the best moment to prevent the dropout of learners in the first two weeks. However, this was only true for the learners in clusters B ("transient learners") and C ("heavy starter learners"). For learners in clusters A ("light learners"), D ("moderate and persistent learners"), and E ("entropic learners"), learner attrition occurred at a different stage. The five clusters of learners' characteristics indicate that course instructors and instructional designers need to design course activities based on diverse learners' traits at different stages of MOOCs. Therefore, MOOC instructors should prevent learners dropping out through various temporal interventions. For example, early intervention could be provided before week 3 to help those “transient learners” and “heavy starter learners” stay engaged in meaningful forum discussions. Instructors should also design more discussion activities in following weeks for those “light learners”, “moderate and persistent learners” and “entropic learners” to facilitate interaction and promote learner engagement.

Moreover, we found that learners’ course grade was associated with their longitudinal pattern of meaningful forum participation, consistent with Wang et al.’s (2015) findings that learners’ active and constructive discussion behaviors predicted their final exam scores. While Wang et al. (2015) did not analyze the discussion behavior over time, in our study, we considered learners’ longitudinal patterns in predicting their summative grades. A temporal dimension allowed us to take a more granular look at learner behavior than the lens of summative counting learners’ total number of posts in a course, echoing Tang et al. (2018; 2019). For example, in our study, “transient learners”, “heavy starter learners” and “moderate and persistent learners” may have the same total number of posts in this course. However, the results showed that the “moderate and persistent learners” are more likely to receive a higher grade because they had more consistent meaningful forum activities. In addition, Tang et al. (2018, 2019) indicated that learners’ temporal dimension of forum participation influenced their course performance. The finding extends evidence from prior studies with a focus on learners’ forum activities of posting topic-relevant content and further reinforces the significance of facilitating meaningful forum discussions in MOOCs. Building upon these findings, we recommend MOOC instructors to encourage learners to persistently participate in meaningful forum discussions over time, rather than only considering their total number of forum posts and comments.

Despite the findings, limitations of the study should be noted. First, the model was designed and tested in only a MOOC in one topic from a course platform, which may not be representative for all the MOOCs. Future research might further validate the model in a wider range of MOOCs and topics. To prevent the learner dropouts, we plan to build a model upon learners' meaningful participation patterns in the early stage of the course to predict the five clusters of learners. Furthermore, this research only analyzed the forum participation through whether the posts were topic-related or not. Based on the results of our topic-related posts, other factors such as self-regulated learning strategies in discussion forum could be considered, since it has been proved that self-regulated learning skills have a great impact on learners' success and dropouts (Moreno-Marcos et al. 2020). Moreover, understanding the relationship between learners' unmeaningful or trivial participation and their course retention can provide additional insights for course design and facilitation practice. Unmeaningful participation such as course complaints and system error reports may have an impact on learners' grades and their dropout rates. Lastly, in our study, we only focus on grades as a summative measure of learner performance. Future research may seek validated measures of learner performance.

2.6. Conclusion

This research investigated how MOOC learners' longitudinal trajectory of meaningful forum participation was associated with their course performance and retention. Using natural language processing and machine learning methods, this study presented a model that classified relevant and irrelevant posts in MOOC forums. This work identified five discussion forum participation patterns and their characteristics. The findings show that there is a relationship between leaners' forum behaviors of posting topic-related content and their course performance. The findings provide insightful implications on refining the courses facilitation and forum design in MOOCs and also helping improve learner performance such as their course grade and course retention in MOOCs. Future works could expand our model to fit more MOOCs and topics.

Table 2.1 Examples of topic-related and topic-unrelated posts

Topic-related examples	Topic-unrelated examples
I'm water-coloring for years, started acrylics a while ago and tried other drawing in-between. For me drawing is relaxing, a balance and completely different to my all-day life/work.	How do we access the reading for the first assignment on fantastic artists?
My favorite artists are Bosch, Botticelli, Braque, Canova, De Chirico, Dali, Davis and many others. I like abstract art because it forces the observer to think about the work.	No button in Firefox but it works in Safari on my MacBook
I applaud your use of 3-D art! However, due to the messiness and disorganization, it reminds me of a child's diorama. I wish you would give it another try and make it more garden-like and less like a craft project.	The quizzes really don’t take much time at all. Why not do it anyway?

Table 2.2 Sub-datasets

Datasets	Number of Posts	Number of Posts in Use	Topic-related	Topic-unrelated
Full set	4215	4108	3526	582
Training set	1232	1186	754	432
Test set	443	443	293	150

Table 2.3 Model result

	Accuracy	Kappa	ROC_AUC
Identified model	0.76	0.506	0.83

Table 2.4 Average posts per week

Cluster	Week1	Week2	Week3	Week4	Week5	Week6	Week7
A	1.7	2.3	1.7	0.8	0.7	0.5	0.6
B	8.6	1	0.5	0.3	0.2	0.2	0.2
C	23.5	2.8	1.8	0.7	0.3	0.3	0.5
D	2.9	1.8	4	8	3	1.6	1.3
E	3.8	11.7	8.1	1.5	1	0.5	0.5

Table 2.5 Students number, proportion, average grade and standard deviation

Cluster	Students	Proportion	Average Grade	Standard Deviation
A	575	0.43	76.8	29.4
B	485	0.37	64.66	37.8
C	99	0.07	71.85	33.52
D	86	0.06	89.13	19.03
E	80	0.06	83.46	21.18
Total	1325	1	73.20	32.98

Table 2.6 Chi-square analysis result

Cluster		Dropout		Total
		Yes	No	
A	Count	93	482	575
	Expected Count	117.6	457.4	575.0
	Adjusted Residual	-3.4	3.4	
B	Count	151	334	485
	Expected Count	99.2	385.8	485.0
	Adjusted Residual	7.3	-7.3	
C	Count	19	80	99
	Expected Count	20.2	78.8	99.0
	Adjusted Residual	-.3	.3	
D	Count	3	83	86
	Expected Count	17.6	68.4	86.0
	Adjusted Residual	-4.0	4.0	
E	Count	5	75	80
	Expected Count	16.4	63.6	80.0
	Adjusted Residual	-3.2	3.2	
Total	Count	271	1054	1325
	Expected Count	271.0	1054.0	1325.0

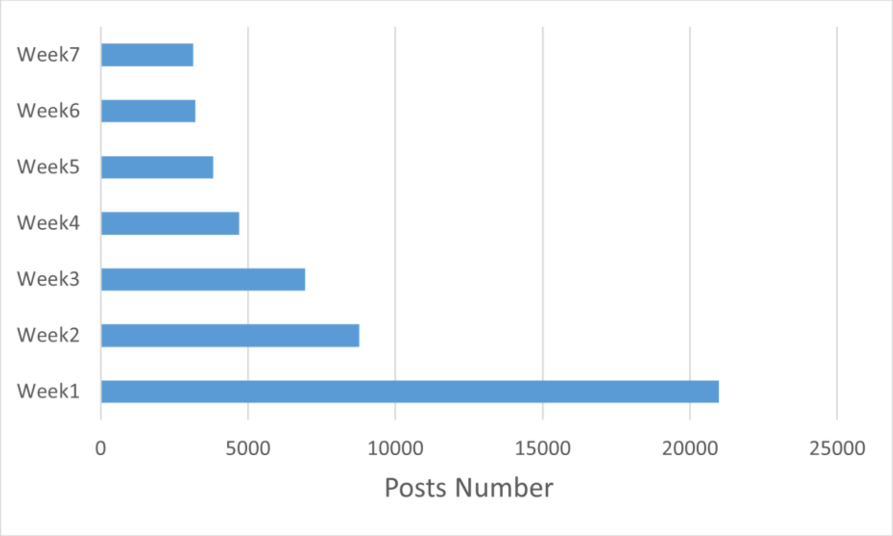


Figure 2.1 Number of posts over seven weeks

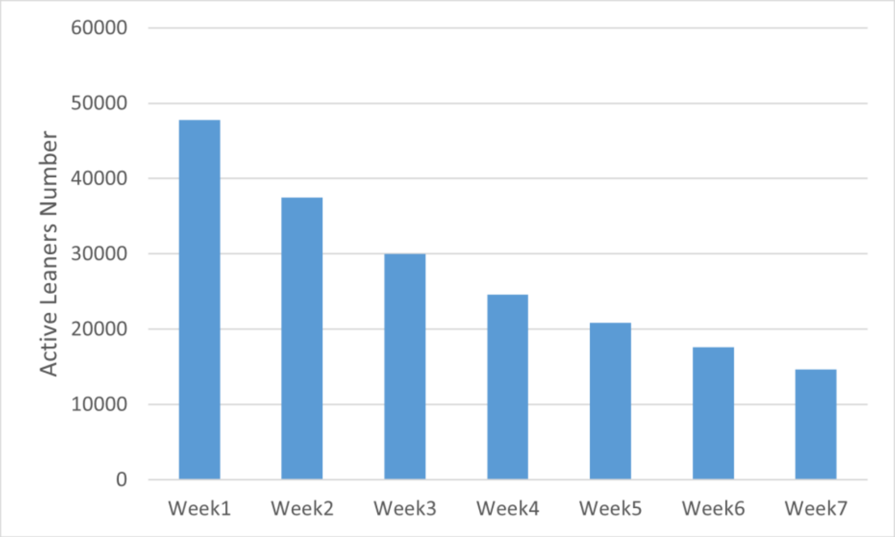


Figure 2.2 Number of active learners over seven weeks

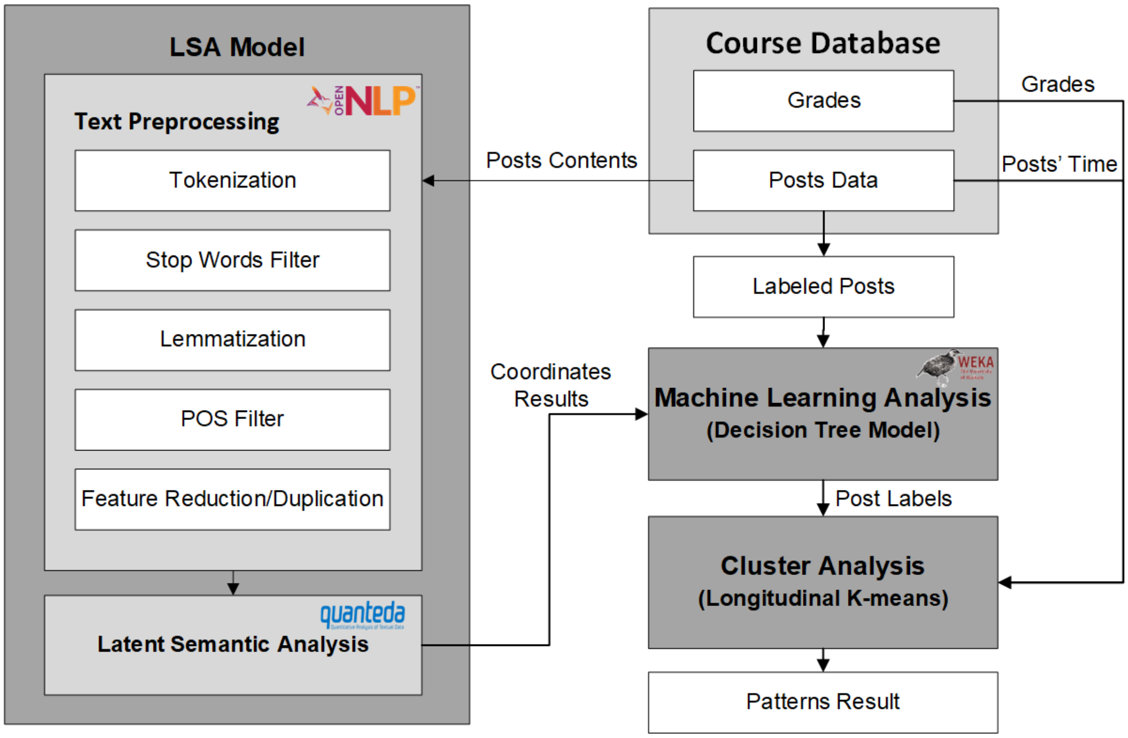


Figure 2.3 System architecture

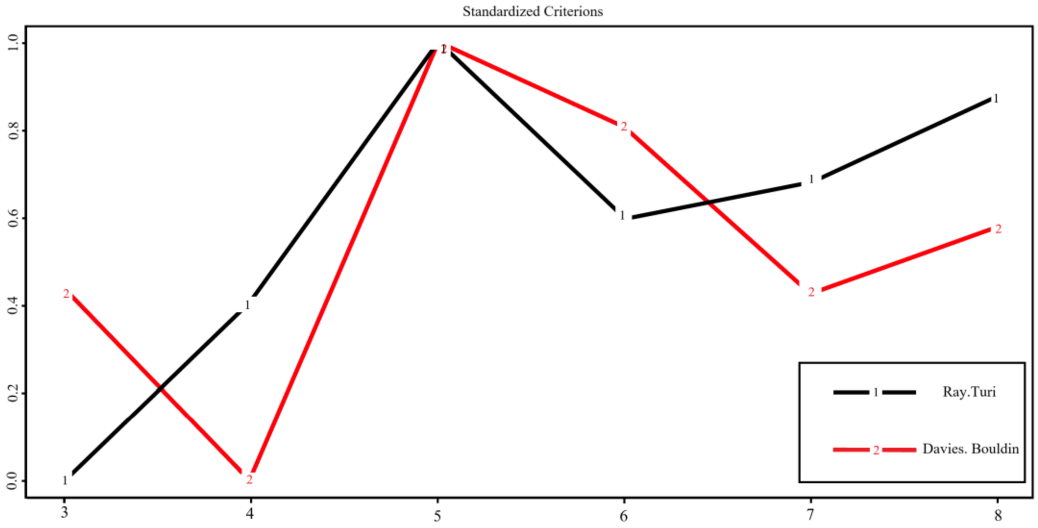


Figure 2.4 Criteria values for varying number of clusters

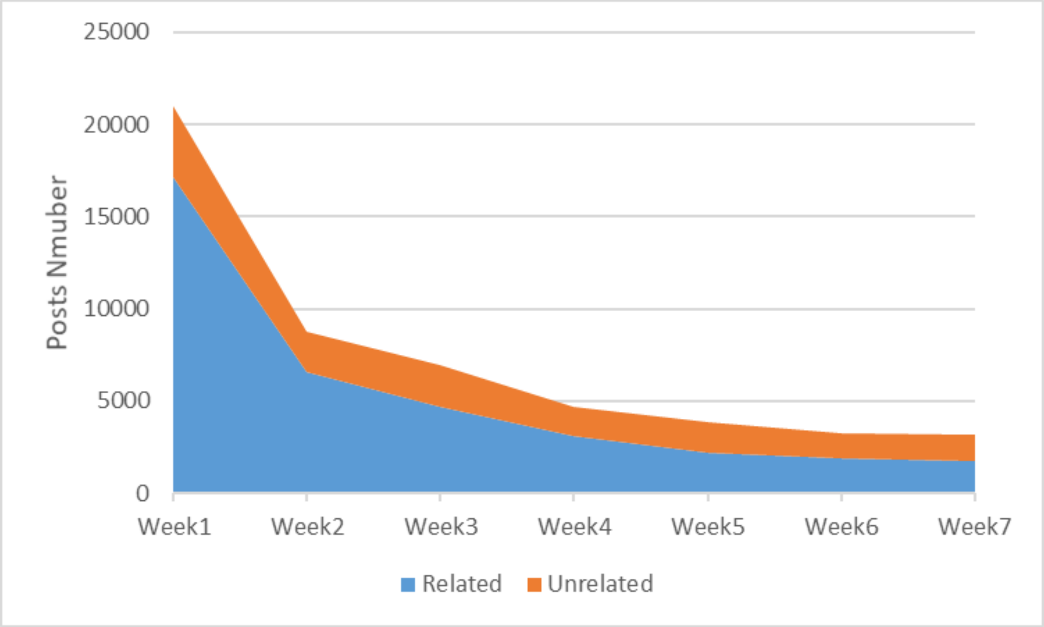


Figure 2.5 Number of topic-related and topic-unrelated posts over time

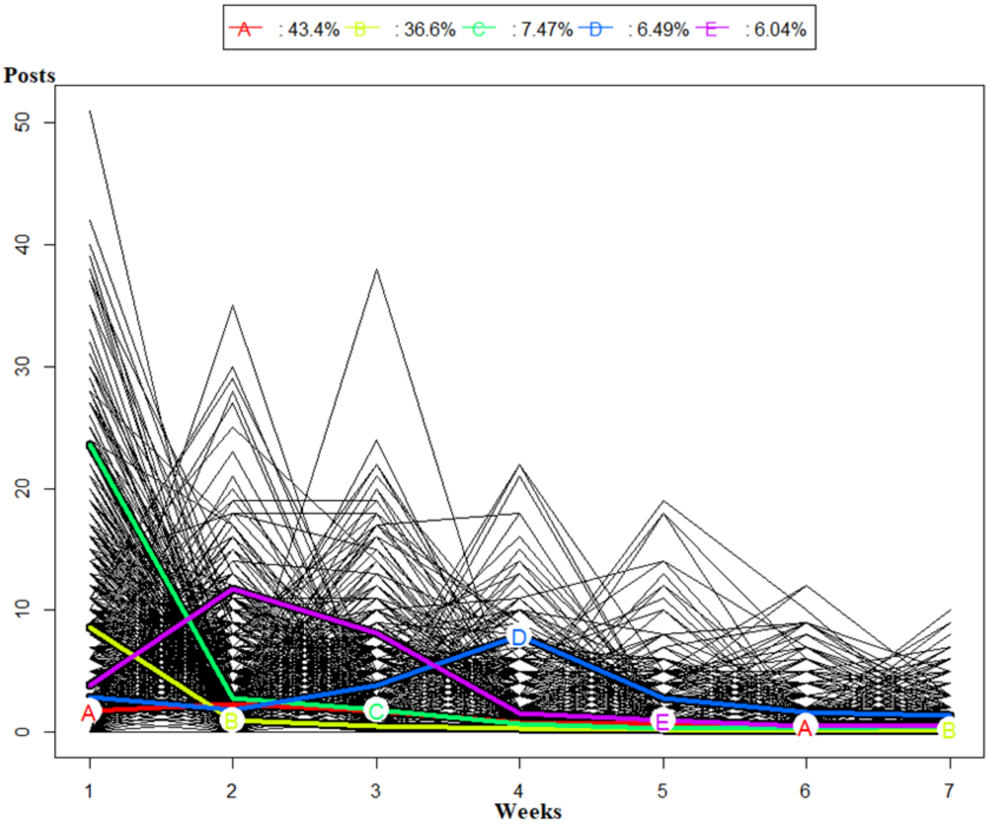


Figure 2.6 Five clusters of learners based on their longitudinal forum participation

CHAPTER 3

ARGUMENTATIVE ESSAY STRUCTURE ANALYSIS

3.1 Introduction

Argumentative writing is an essential part of students' learning (Crossley et al., 2022; Pessoa et al., 2017; Lee & Deakin, 2016). Several educational organizations evaluate students' writing skills in their examinations. Considering the student-teacher ratio and the teacher shortage (Sutcher et al., 2019), the manually evaluation process becomes more and more time and effort consuming (Uto, 2021). With the purpose of avoiding the need for manual grading effort, automatic essay scoring (AES) aims to grade students' essay without human guidance. Most of the previous AES research focused on providing a holistic score (Ke & Ng, 2019). Usually, an AES system takes a student's essay text as an input and assigns a score for the essay (e.g., Alikaniotis et al., 2016; Dong et al., 2017; Süzen et al., 2020; Taghipour & Ng, 2016). As a result, these AES systems can only provide a final score of the essay, which summarizes the quality of an essay as a whole.

Although the AES tools can provide a holistic score for students essays and save essay grading efforts, they are not functionally advanced enough to be applied in classroom settings (Ke & Ng, 2019). Both teachers and students expect feedback from the AES tools. Teachers demand details of the graded essay and students need feedback to improve their essay. Especially, it is essential to provide feedback for students who received relatively low essay scores and explain the reasons for their scores. Thus, lack of feedback becomes a weakness of the existing AES tools.

Some early research in AES applied feature engineering and included features that might influence the essay score in their development of the AES systems (Persing & Ng, 2015; Rahimi et al., 2017; Yang & Zhong, 2021). Several features have been proved to affect the score, for example, the coherence and relevance of the essays (Yang & Zhong, 2021), the effective use of evidence (Rahimi et al., 2017), the strength of the arguments (Persing & Ng, 2015) and so on. However, there are still some possible influential features that remain unexplored. Assessing an essay by including all features that might affect its score is still a big challenge. In addition, most of these studies have been limited to applying feature engineering to predict a holistic score for the essay. The assigned score of the essay still cannot be fully explained.

Argumentative structure is one of the dominant features that could influence the score of an argumentative essay (Persing & Ng, 2015). The past decade has seen the rapid development of argument mining (Lawrence & Reed, 2020). Argument mining has been applied to identify the claims and premises in argumentative essays (e.g., Wan et al., 2021; Chakrabarty et al., 2019; Persing & Ng, 2016; Afrin et al., 2021). With the ability to automatic identify and extract the argumentative structure of an essay, argument mining has been applied to education research (Cabrio & Villata, 2018). Recent evidence suggests that argument mining could be used to support AES tools (Nguyen & Litman, 2018). So far, however, there has been little discussion about what kind of argumentative structure contributes to the score and the research has tended to focus on predicting a score rather than providing feedback.

To date, only a limited number of researchers have attempted to provide essay feedback for students (Afrin et al., 2021; Nagata, 2019; Ye & Manoharan, 2019; Zhang et al., 2019) and none of them focused on providing essay argumentative structure feedback. This paper aims to understand the association between essays' argumentative structures and scores and provide a feedback tool which can support the automatic essay scoring system. To extract the argumentative elements of an essay, we applied an argument mining approach and proposed a cross-prompt, sentence-level ensemble model to classify the argumentative elements. In addition, a sequential pattern mining process was applied to extract the argumentative structures of the previous essays. In order to provide solid feedback, we seek to understand what kinds of structure contribute to the score. Thus, in this paper, we address the following research questions:

RQ1: How reliable are the argument elements classification models when applied to datasets with different prompts?

RQ2: What is the most frequently used argumentative structure from the essay dataset?

RQ3: Do essays with the same score share similar lengths and structures?

3.2 Literature Review

3.2.1 Automatic Essay Scoring

The early AES studies applied feature engineering and traditional machine learning algorithms to provide a score for the input essay (e.g., Dascalu et al., 2017; Persing & Ng, 2015; Rahimi et al., 2017; Zhang et al., 2019). For example, Rahimi et al. (2017) proposed a score prediction method based on students' effective use of evidence and their organization of ideas and evidence to support their claims. They provided several features with rubric to predict the score of the evidence and organization dimensions. In their research, a random forest model was used as the classifier for scores. However, they identified evidence through a simple word-matching algorithm with a manually provided list of keywords for each topic. In a following study, Zhang et al. (2019) further applied this evidence rubric and provided feedback regarding students’ use of evidence. Persing and Ng (2015) attempted a sentence-level, support vector machine (SVM) based AES model to score the essays through an argument strength dimension. The model considered more than ten features, including the number of claims and supports, transitional phrases, coreference and so on. However, their argumentative elements identification process was still based on a string-matching approach with a sentence labeling rule. Recently, researchers have shown an increased interest in applying deep learning models to predict essay score (Alikaniotis et al., 2016; Dong et al., 2017; Taghipour & Ng, 2016; Uto et al., 2020). Taghipour and Ng (2016) presented a recurrent neural networks approach to learn the relationship between essays and their scores. The models were built without feature engineering and provided a holistic score based on the input text. In a similar work, Dong et al. (2017) applied a recurrent convolutional neural network to learn text representation and provided a holistic score for the input essay. Although deep learning approaches can learn the relationship between essay text and score without feature engineering, this kind of AES tools provide limited information for classroom teachers and students.

Recent years have witnessed the development of natural language processing (NLP), especially, the development of deep learning massive language models. The deep learning massive language models like BERT (Bidirectional Encoder Representations from

Transformers) (Devlin et al., 2018) and GPT (Generative Pre-Training) (Brown et al., 2020) have achieved state-of-the-art results in many NLP tasks. Thus, the research focus of AES research community moves from traditional machine learning and early deep learning approach to transformer-based models, in particular, the transformer architecture BERT. Yang et al. (2020) developed a BERT based AES model and compared the model with previous deep learning-based AES models. Their results showed that the pre-trained language model (BERT) outperforms the previous state-of-the-art neural models. In contrast, Mayfield and Black (2020) argued the necessity of applying Transformer models, since the state-of-the-art accuracy came with significant tradeoffs. With the help of pre-trained BERT, Yang and Zhong (2021) proposed a hierarchical structured model to extract semantic features at both sentence-level and document-level. The extracted semantic features were then used to evaluate coherence and relevance in the essays and compute the final score. In a recent study, a hierarchical BERT-based transfer learning approach was carried out by Xue et al. (2021). Instead of providing a single holistic score for the essay, the model also provided scores for different features of the essay including grammar, lexicon, idea supporting and so on. Nonetheless, with the long-term goal of producing feedback for students and teachers, the model cannot provide details on why the essay received a certain score under each category. Although BERT based AES models can achieve better performance on score prediction without feature engineering, it is still necessary for the novel AES systems to consider features that could influence the final score of an essay.

3.2.2. Argument Mining

Argument mining is a research area within NLP, aiming at extracting and identifying argumentative elements and structures from text (Cabrio & Villata, 2018). Stab and Gurevych (2014) proposed a SVM based approach and classified sentences into four categories: MajorClaim, Claim, Premise and None. Their results showed that the model can obtain an F1-score of 0.726 for identifying argumentative components. In a following study, Nguyen and Litman (2018) improved the argument component identification model developed by Stab and Gurevych (2014). They used some argumentative features like number and fraction of argument components over total number of sentences, number of claims, number of premises and so on to improve the AES systems. Their results showed that the argument mining could improve the performance of the holistic score-based AES systems. Deep learning massive language models have also been applied into argument mining (Alhindi & Ghosh, 2021; Niculae et al., 2017; Wang et al., 2020). Wang and his colleagues (2020) adopted BERT to mine three argumentation components: major claims, claim and premises. Their approach considered both essay-level, paragraph-level and word-level classification. The overall F1 score of their model is 0.64. In another study, Alhindi and Ghosh (2021) applied a token-level classification to identify claim and premise token of an essay. The overall F1 score of their result is 0.57. During their experiments, they noticed that multitask models can identify some instances missed by the single task model. In addition, they claimed that some sentences may contain multiple claims which caused misclassification.

It has been proved that the performance of the AES model can be improved by applying the argument mining (Nguyen & Litman, 2018). However, previous research still focused on predicting a final score for the input essay instead of providing feedback. As a result, the implementations of previous AES tools do not support feedback providing. It is essential to provide feedback and the reasons behind a score. In this study, we aim to build an essay argumentative structure feedback tool which can support the current AES systems.

3.3 Data

3.3.1 Feedback Prize Dataset

To train the argumentative elements classification model, we used the dataset of the Kaggle competition "Feedback Prize - Evaluating Student Writing" provided by Crossley et al (2022). The dataset consisted of 15,594 argumentative students' essays from about 15 prompts, written by U.S students in grades 6-12. The essays were annotated for seven commonly used elements, including Lead, Position, Claim, Counterclaim, Rebuttal, Evidence and Concluding Statements. The explanations of each label are shown in Table 3.1. The overall inter-rater reliability of the dataset was .73.

According to the explanations on the competition webpage (as shown in Table 3.1), counterclaim and rebuttal are also claims. In addition, as mentioned by the competition host, counterclaims and rebuttals had the lowest reliability. These two elements were often labeled as claims. Thus, in our experiments, we merged Claim, Counterclaim and Rebuttal into a single label Claim. After some investigation, we found that there were overlaps between the labels. That is to say, the data labeled as Lead and Concluding Statements consisted of positions, claims, and evidence. Since we applied a sentence-level prediction and the prediction process did not consider the position of the sentence, we decided to exclude the data labeled as Lead and Concluding Statements from the dataset. After preprocessing the Feedback Dataset, we applied three labels: Position, Claim, Evidence.

3.3.2 ACT Previous Test Essays Dataset

To further analyze the association between essay structure and score, we obtained a large number of essays from the previous ACT tests. The previous ACT tests essays contained 13,990 essays from 27 prompts, collected from the previous ACT writing tests with dates ranging from September 2020 to March 2021. Students who took the tests come from more than 50 countries and over 500 regions. To test the generality of the model, we also applied a testing set which contained 30 essays with 723 sentences from the previous ACT writing tests.

3.4. System Design

In this section, we describe the datasets used to train and test the model and the proposed argumentative essay structure feedback providing tool. The system contained two parts: the ensemble model block and the essay analysis block. The system overview is shown in Figure 3.1. First, we developed an ensemble model to classify the argumentative elements based on several pre-trained deep learning massive language models. To provide feedback for future students' essays, we need to understand the associations among the use of argumentative elements, the use of argumentative structure and essay scores in the previous essays. Thus, we further applied the ensemble model to classify the argumentative elements and extract argumentative structures from the previous ACT tests essays.

3.4.1 Datasets

The datasets we used to train and test the ensemble model are shown in Table 3.2. After preprocessing the Feedback Dataset, we obtained 120,630 data with three labels: Position, Claim and Evidence. Note that each instance of the data in the Feedback Dataset can contain multiple sentences. We randomly split more than 100K data into the training set and the validation set of the models. To improve the performance of the ensemble model, we further increased the size of the training sets for models used to predict position and evidence. The testing set from Feedback Prize dataset contained 10K randomly selected data. To test the generality of the model, 723 sentences (30 essays) were annotated by three researchers, two of them are experts on argumentative essay grading. In total, we built two training sets, two validation sets and two testing sets to train and test the ensemble model.

3.4.2 Ensemble Model Block

We first built a multiclass classification model to classify all three kinds of argumentative elements. The model was built based on the pretrained model: DeBERTa (He et al., 2020). DeBERTa is a transformer-based language model which improves BERT (Devlin et al., 2018) with two novel techniques: the disentangled attention mechanism and the enhanced mask decoder (He et al., 2020). DeBERTa was pretrained with 80GB training data and achieved better performance on the majority of natural language understanding tasks.

We then built four binary classification models to classify positions and evidence based on DeBERTa and DistilBERT (Sanh et al., 2019). DistilBERT is another transformer-based language model which is a distilled version of the BERT model. DisilBERT reduces the size of a BERT model but retains 97% of its language understanding capabilities (Sanh et al., 2019). We included DistilBERT as another language model because it was trained on a different corpus which could provide more information on the sentences.

3.4.2.1 The Voting Scheme

To synthesize the results of five different models, a voting scheme was needed to decide the final label of a sentence. We emphasize that our goal is to find as many positions and evidence as we can, the algorithm of the voting scheme was designed as follows (Algorithm 3.1).

3.4.2.2 Evaluation Metrics

To evaluate the models, we proposed the following evaluation scheme: accuracy, Macro-F1 and recall. The goal of this research is to find the associations among essay elements, structures used in the essay and the score the essay received. We tried to find as much positions and evidence as we can. Thus, we included recall value as an evaluation metric.

3.4.3 Essay Analysis Block

After applying the ensemble classification model on the ACT previous test essays, we could further analyze the elements and structures used in the previous essays. We first analyzed the association between essay length and score. Then we assigned a label of each sentence from the previous essays and calculated the average number of positions, evidence used and their proportions. Finally, to extract the argumentative structures of the previous essays, we applied a sequential pattern mining process on the label sequences of the previous essays.

3.4.3.1 Sequential Pattern Mining Process

The sequential pattern mining process was done by the TKS (top-k sequential pattern mining) (Fournier-Viger et al., 2013) algorithm in SPMF (Fournier-Viger et al., 2016). The TKS algorithm was designed to find the frequent subsequences in a sequence database. The advantages of applying TKS algorithm on this task is that it provides the interfaces to specify the minimum and maximum length of the result pattern. In addition, some required items can be chosen when applying the algorithm. As a result, the chosen items must appear in every pattern found. Finally, the algorithm also allows the user to set a number of gaps between two consecutive itemsets in a pattern.

According to the average length of the essays and the proportion of the positions and evidence, for essays received score between 7-12, the length of the result sequences was set to 8 to 12. For essays receiving scores from 2 to 6, we set the length of the result sequences to 5 to 9. The gap between each itemset was set to 3, in order to cover the whole essay as much as possible. In addition, to better extract the argumentative structure, the label position and evidence were chosen to be the required items for each subsequence.

3.4.3.2 Feedback Providing Process

Finally, we provided essay structure feedback based on the knowledge of the previous ACT tests essays and the information we extracted from the current essay. Feedback was provided based on the length, argumentative elements usage, and argumentative structure used in the current essay. In total, there were 17 different cases of feedback (as shown in Table 3.3). After applying the ensemble model on the sentences of the input essay, a label of position, claim or evidence was assigned for each sentence. First, some suggestions on how to include more positions and evidence were provided based on the length and argumentative elements usage of the input essay. Then, the result sequence of the labels was compared with the top subsequences of the previous essays to extract the essay structures. In addition, considering the position of the labels in the sequence, a position or evidence can be further considered as a lead or conclusion of the essay. Finally, essay structure suggestions were provided according to the structure, lead, conclusion information extracted from the input essay and the scoring rubric.

3.5 Results

3.5.1 Ensemble Model Results

We present our experiment results of the ensemble model in this section. We employed three evaluation metrics: accuracy, recall value and F1 value. The results of the models are shown in Table 3.4. As we can see from the results, the triple classification model achieved better overall accuracy and F1 but lower recall on the Feedback Prize testing set. The ensemble model scored better on almost all the evaluation metrics on the ACT testing set.

3.5.2 Previous ACT Tests Essays Analysis Results

First, we analyzed the previous essays through their length. As can be seen from Figure 3.2 (a), most of the previous essays were between 8 and 35 in length. After that, we analyzed the length, number of positions used, number of claims used, and number of evidence used for essays with different scores. The results can be found in Table 3.5 and Figure 3.2 (b). From our results, longer essays tended to receive higher scores. In addition, higher scored essays always contained more positions and evidence.

The sequential pattern mining results showed that the most frequently used structure in the previous essays was position-claims-evidence-claims (PCEC). Around 40% of students had this structure in their essay. This result also indicated that a huge number of the students did not include a position or evidence in their essays and their essays did not have an organizational structure.

Finally, we performed a sequential pattern mining process based on the classification results of the previous essays with different scores. The results of the sequential pattern mining procedure (top three sequences) are shown in Table 3.6. We note that there are overlaps between different mining results. As we can see from the sequential pattern mining results, essays with score 10 to 12 used a different structure which was evidence-claims-position (ECP). All the essays with lower scores tended to be the position-claims-evidence-claims (PCEC) structure. Besides, the lower the essay scored, the higher the possibility that the essay tended to have no structure. For example, for those essays scored less than 4, only 3% of them had an organizational structure. In contrast, approximately 70% of the essays which received a score of 12 used evidence-claims-position structure.

3.5.3 The Feedback Proving Process

Based on the knowledge of the ACT tests essays, we built a feedback tool and provided feedback to students. Figure 3.3 shows a prototype of the tool we designed. The tool can provide suggestions on three dimensions: length, argumentative elements, and essay structures. The detailed feedback for the input essay included whether a lead or conclusion needs to be included, if there is a clear position used in the essay, the number of evidence used and so on. If no organizational structure was detected, the tool would also recommend the two frequently used argumentative structures (i.e., PCEC and ECP) and explain the advantages of applying an argumentative structure in an essay.

3.6 Discussion

The purpose of the current study is to provide an argumentative structure feedback tool for students' essays. This paper applied an NLP-based ensemble model to classify argumentative elements and extract argumentative structures. The ensemble model achieved high overall accuracy on both datasets with different prompts. The significance of this paper is that it highlighted the associations among argumentative elements, structures and essay scores. Our study demonstrated that certain argumentative structures could contribute to the score. Based on the analysis results, we further built a feedback tool to better support classroom teaching and the current AES tools.

The most obvious finding emerged from the analysis was that longer essays tended to receive higher scores, which was consistent with the finding of (Persing & Ng, 2015) that a short essay had less potential to make a strong argument. In addition, our finding provided evidence that the more positions and evidence used in an essay, the higher the score, which reflected the findings of (Persing & Ng, 2015; Rahimi et al., 2017). This is also a possible explanation for why longer essays get higher scores. These associations suggested that including more positions and evidence to support the claims and positions can facilitate a stronger argumentative essay.

In previous studies, several features have been proved to affect essay score, including coherence, relevance, and the strength of the arguments (Persing & Ng, 2015; Yang & Zhong, 2021). Our results demonstrated that some certain kinds of argumentative structure can have a positive impact on essay score. The first research question in this study sought to determine the most frequently used structure in the previous ACT tests essays. The sequential pattern mining results suggested that the most frequently used structure was position-claims-evidence-claims (PCEC). This is a normal structure that starts with a position, followed by claims and evidence to support this position. Based on the results, only around 40% of essays applied this structure, indicating that a large percentage of essays did not apply an organizational structure. With respect to the second research question about essay length and structure, the sequential pattern mining results of essays with different scores indicated that the PCEC structure was frequently used in essays with scores between 7 and 9. According to the ACT essay scoring rubric, the essays within this score range are usually considered as exhibiting a basic organizational structure.

The most striking result emerged from the data was that it pinpointed another skillful organizational strategy used by essays with higher scores (10 to 12). These essays tended to use an evidence-claims-positions (ECP) structure. Surprisingly, approximately 70% of the essays with a score of 12 applied the ECP structure. After manually inspecting some samples of the essays with this structure, we found that these essays usually started with stories or facts. For example, a story of the author was used as a lead and the essay was concluded with a position. For the essays received lower scores (i.e., score 2 to 3), as illustrated in Table 3.6, only around 3% essays applied the position-claims-evidence structure. For the essays received scores between 4 and 6, less than 30% of them had an organizational structure. These findings suggested that the organizational structure is one of the dominant features that could influence the score of an argumentative essay. Especially, no organizational structure is the mark of receiving a lower score. These results were partially consistent with other research which found that the argumentation structure can be used to improve the performance of AES (Nguyen & Litman, 2018). Based on these results, it can thus be concluded that lower scored essays were usually short of length, lacked position and evidence, and did not apply an organizational argumentative structure. Further feedback and help could be provided for the essays that fell into this category.

In our feedback tool, we provided feedback on students' use of argumentative structure. Previous works have proposed feedback tools for student essays through evidence usage (e.g., Afrin et al., 2021; Zhang et al., 2019), preposition usage (e.g., Nagata, 2019), and semantic meaning (e.g., Ye & Manoharan, 2019). Our work provided further support on argumentative elements and structures in the essay. The feedback tool provided detailed information for teachers on the argumentative elements used in the input essay. For the essay short in length, the feedback focused on adding more positions and evidence to support the arguments. Based on the students' structure use, we recommended some structures for students with lower scores, aiming to help them make stronger arguments and improve their scores. With the help of our feedback tool, the existing AES tools could be improved by adding detailed feedback on students' argumentative elements and structures usage.

To identify the argumentative elements and extract the argumentative structure of the essays, this study proposed a cross-prompt, ensemble learning model. According to Table 3.4, all model variations were able to learn the tasks and perform accurate classification results on the Feedback Prize testing set. However, for the ACT testing set, the ensemble model outperformed the base models by a large margin. The ensemble model was able to find nearly 20% more positions and evidence compared to the base models. This might be caused by the different training corpus of the pre-trained language models. These results further confirmed the ability of the ensemble model to work across multiple datasets and achieve sufficient accuracy. Although the ensemble model achieved an overall accuracy of .835 on the ACT testing set, the results of recall and F1 for positions and evidence did not perform as good as the results on the Feedback Prize dataset. This inconsistency might be caused by the cross-prompt implementation. We believe the performance of the ensemble can be increased by adding more essay sentences from the targeted prompts. In addition, the declined model performance for positions and evidence may also be caused by the sentence-level approach. After inspecting some misclassified cases, we found that our experts also have disagreement on the labels of these cases. Sometimes, an extremely long sentence may contain multiple argumentative components. Although our ensemble learning approach overcame this weakness to a certain degree, a border detection approach could be considered for follow-up studies.

This study has some implications for future research. Despite the fact that our ensemble model achieved a good enough overall accuracy on different datasets, for future implementations, a border detection approach could be considered to test if the performance of the models can be improved. In addition, the classification models for argumentative elements can be trained with the targeted prompt datasets. This study only provides feedback on the use of argumentative elements and structures. Several features remain unstudied at present, for example, the number of perspectives on the given prompt, the transitions between and within paragraphs, the relevance of the examples to the prompt and so on. Further studies, which take these variables into account, will need to be undertaken. Moreover, further investigations and experiments on how these features could improve the massive languages model-based AES models are strongly recommended.

3.7 Conclusion

In this study, we present a methodology to extract the argumentative structures in students' essays and propose an argumentative structure feedback tool to support the current AES systems. We used a deep learning massive language-based ensemble model to identify the argumentative elements in essays and achieved high performance compared to previous works. The findings of this study highlight that the use of argumentative structure is a significant factor in assigning scores for essays. The proposed feedback tool could assist classroom teachers to easily extract argumentative structures from students' essays and help students make stronger arguments and improve their scores.

ALGORITHM 3.1 Voting Scheme Algorithm

```
Input: A vector of models' prediction results on the sentence prediction_results
Output: Label of the sentence
position, evidence, claim = 0
for each predict_label in prediction_ results
    if predict_label == 'position'
        position += 1
    else if predict_label == 'evidence'
        evidence +=1
    else
        claim +=1
    end if
end for

if max (position, evidence, claim) == position
    return 'position'
else if max (position, evidence, claim) == evidence
    return 'evidence'
else
    return 'claim'
end if
```

Table 3.1 Labels and explanations [a]

Label	Explanation
Lead	An introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis.
Position	An opinion or conclusion on the main question.
Claim	A claim that supports the position.
Counterclaim	A claim that refutes another claim or gives an opposing reason to the position.
Rebuttal	A claim that refutes a counterclaim.
Evidence	Ideas or examples that support claims, counterclaims, or rebuttals.
Concluding Statements	A concluding statement that restates the claims.

[a] The explanations were taken directly from the competition webpage: https://www.kaggle.com/competitions/feedback-prize-2021/data

Table 3.2 Datasets

Model	Training Set	Validation Set	Feedback Testing Set	ACT Testing Set
deberta-base-3 labels	100630	10000	10000	723
deberta-base-position	108630	2000	10000	723
distilbert-base-position	108630	2000	10000	723
deberta-base-evidence	108630	2000	10000	723
distilbert-base-evidence	108630	2000	10000	723
ensemble model	NA	NA	10000	723

Table 3.3: Feedback cases

Category	Cases
Length	1. The length of the essay is less than 20
	2. The length of the essay is between 20 and 30
	3. The length of the essay is more than 30
Position Used	4. No position used.
	5. Position’s proportion is less or equal than .05
	6. Position’s proportion is larger than .05
Evidence Used	7. No evidence used.
	8. Evidence’s proportion is less or equal than .17
	9. Evidence’s proportion is larger than .17
Lead	10. No clear lead sentences
	11. Positions used as lead sentences
	12. Evidence used as lead sentences
Conclusion	13. No clear conclusion sentences
	14. Positions used as conclusion
Argumentative Structure	15. No clear argumentative structure used
	16. Position-Claims-Evidence structure applied
	17. Evidence-Claims-Position structure applied

Table 3.4 Model evaluation results[a]

	Feedback Prize								*Affiliation* Writing							
Models	Accuracy/F1		Recall/F1 (Position)		Recall/F1 (Evidence)		Recall/F1 (Claim)		Accuracy/F1		Recall/F1 (Position)		Recall/F1 (Evidence)		Recall/F1 (Claim)	
Position*	**0.943**	**0.877**	**0.811**	**0.786**	N/A		N/A		**0.936**	**0.736**	**0.351**	**0.465**	N/A		N/A	
Position^	0.94	0.864	0.747	0.762	N/A		N/A		0.935	0.728	0.333	0.447	N/A		N/A	
Evidence*	**0.913**	**0.906**	N/A		**0.862**	**0.881**	N/A		**0.881**	**0.791**	N/A		**0.521**	**0.629**	N/A	
Evidence^	0.909	0.901	N/A		0.85	0.874	N/A		0.871	0.773	N/A		0.5	0.601	N/A	
3 labels*	**0.865**	**0.848**	0.778	**0.792**	0.87	**0.879**	**0.884**	**0.874**	0.819	0.679	0.351	0.476	0.493	0.603	**0.956**	0.888
Ensemble	0.854	0.838	**0.836**	0.773	**0.895**	0.878	0.827	0.858	**0.835**	**0.73**	**0.544**	**0.59**	**0.679**	**0.7**	0.909	**0.893**

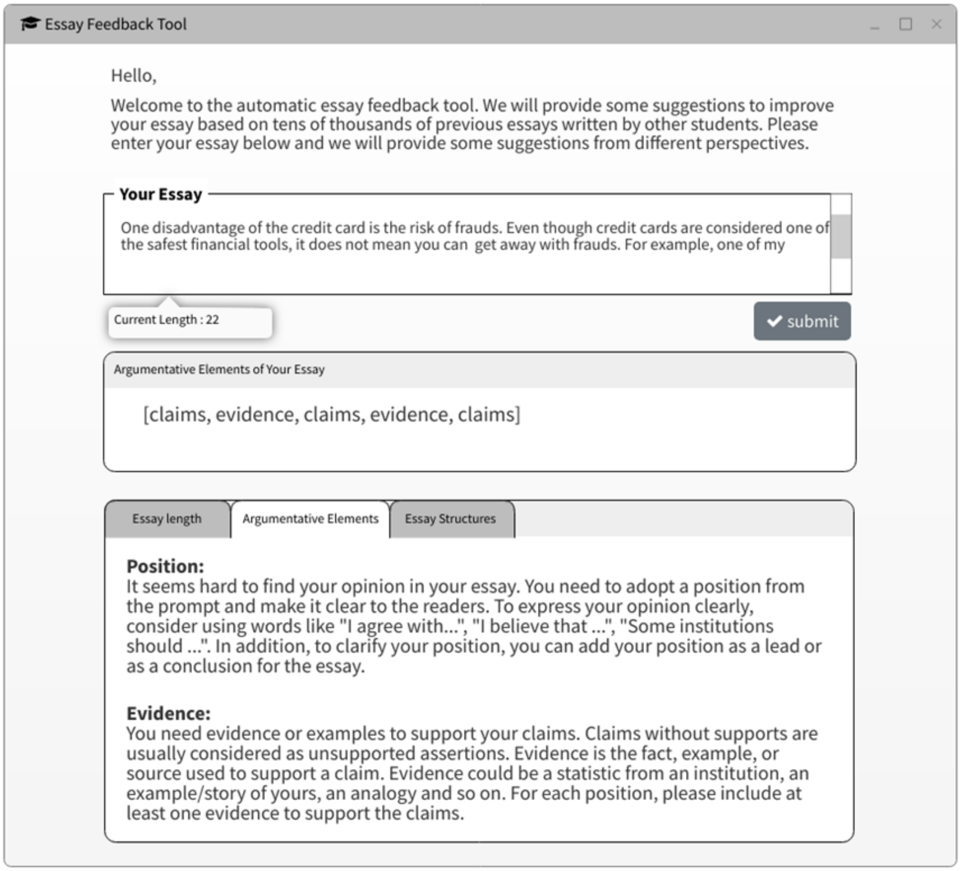
(The highest accuracy/recall/F1 are highlighted in bold. *: DeBERTa-base, ^: DistilBERT-base models)

Table 3.5 Essays data analysis with different scores

Score	Number of Essays	Average Sentences	Average Positions	Average Claims	Average Evidence
2	215	4.237	0.237	2.73	1.27
3	221	6.466	0.443	4.33	1.692
4	805	10.375	0.758	6.896	2.722
5	1018	12.83	0.836	9.1	2.898
6	2929	17.822	1.092	12.955	3.774
7	2094	21.538	1.267	16.127	4.144
8	4062	26.2	1.591	19.839	4.772
9	1228	30.629	1.8	23.368	5.46
10	1032	32.99	2.059	25.095	5.836
11	302	36.662	2.172	27.76	6.728
12	84	40.976	2.988	30.44	7.548

Table 3.6 Sequential pattern mining results

Score	Top Three Sequences	Counts and Proportion
Score 12	evidence | claim | claim | claim | claim | claim | claim | position	59 (70.24%)
	claim | evidence | claim | claim | claim | claim | claim | position	58 (69.05%)
	evidence | claim | claim | claim | claim | claim | position | claim	58 (69.05%)
Score 10-12	claim | evidence | claim | claim | claim | claim | claim | position	807 (56.91%)
	claim | claim | evidence | claim | claim | claim | claim | position	796 (56.14%)
	evidence | claim | claim | claim | claim | claim | claim | position	786 (55.43%)
Score 7-9	position | claim | claim | claim | claim | evidence | claim | claim	2982 (40.38%)
	position | claim | claim | claim | evidence | claim | claim | claim	2978 (40.33%)
	position | claim | claim | claim | claim | claim | evidence | claim	2938 (39.79%)
Score 4-6	position | claim | claim | claim | evidence	1348 (28.37%)
	position | claim | claim | evidence | claim	1344 (28.28%)
	position | claim | evidence | claim | claim	1318 (27.74%)
Score 2-3	position | claim | claim | claim | evidence	14 (3.21%)
	position | claim | claim | evidence | claim	14 (3.21%)
	position | claim | evidence | claim | claim	13 (2.98%)



Hello,

Welcome to the automatic essay feedback tool. We will provide some suggestions to improve your essay based on tens of thousands of previous essays written by other students. Please enter your essay below and we will provide some suggestions from different perspectives.

Your Essay

One disadvantage of the credit card is the risk of frauds. Even though credit cards are considered one of the safest financial tools, it does not mean you can get away with frauds. For example, one of my

Current Length : 22

Argumentative Elements of Your Essay

[claims, evidence, claims, evidence, claims]

Position:
It seems hard to find your opinion in your essay. You need to adopt a position from the prompt and make it clear to the readers. To express your opinion clearly, consider using words like "I agree with...", "I believe that ...", "Some institutions should ...". In addition, to clarify your position, you can add your position as a lead or as a conclusion for the essay.

Evidence:
You need evidence or examples to support your claims. Claims without supports are usually considered as unsupported assertions. Evidence is the fact, example, or source used to support a claim. Evidence could be a statistic from an institution, an example/story of yours, an analogy and so on. For each position, please include at least one evidence to support the claims.

Figure 3.3 The prototype of the feedback tool

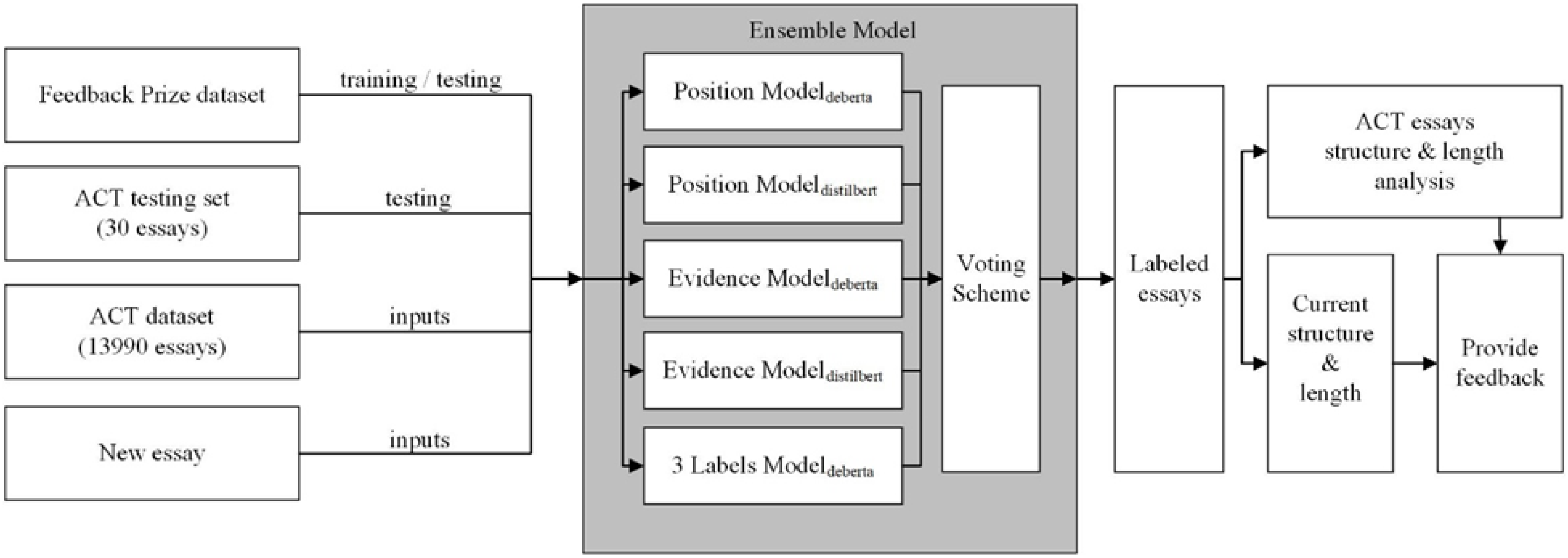
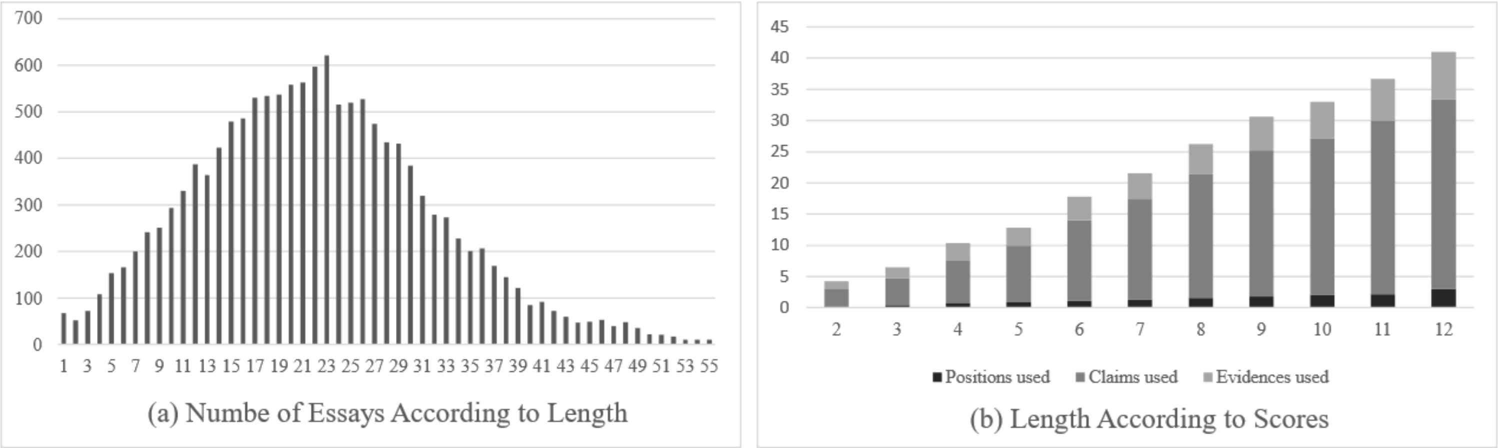

Feedback Prize dataset
training / testing
ACT testing set (30 essays)
testing
ACT dataset (13990 essays)
inputs
New essay
inputs
Ensemble Model
Position Model$_{deberta}$
Position Model$_{distilbert}$
Evidence Model$_{deberta}$
Evidence Model$_{distilbert}$
3 Labels Model$_{deberta}$
Voting Scheme
Labeled essays
ACT essays structure & length analysis
Current structure & length
Provide feedback

Figure 3.1 System overview



(a) Numbe of Essays According to Length

(b) Length According to Scores

Figure 3.2 Length analysis results

CHAPTER 4

MOOC DROPOUT PREDICTION WITH FORUM PARTICIPATION

4.1 Introduction

Massive Open Online Course (MOOC) is an education method that has emerged in the past decade, offering new solutions for distance education, life-long learning, and resource sharing. MOOCs are popular because of their open nature. The unlimited enrolled number, the non-requirement on education level, and the absence of geographical limitation attract learners from all over the world to participate in MOOCs. The discussion forums of MOOCs are primary venues for learners and instructors to interact with each other in MOOCs (Huang et al., 2014; Wise et al., 2017). It provides a learning community for learners and help learners construct collaborative knowledge (Boroujeni et al., 2017; Galikyan et al., 2021). However, with a high enrollment number, both the dropout rate of over 90% and the low cohesiveness of the learning community become critical problems of MOOCs (Boroujeni et al., 2017; Eriksson et al., 2017; Goopio & Cheung, 2021). With unlimited enrollments, MOOC forums are massively filled with a great amount of unstructured information. The high ratio of learner-to-instructor environment results in chaos, information overload, and limited support in MOOC discussion forum All of this leads to low satisfaction and high drop rates (Alrajhi et al., 2020; Capuano et al., 2021; Wise et al., 2017). To solve these challenges, researchers have sought to understand the association between learners' forum behavior and course completion (Goel & Goyal, 2020; Handoko et al., 2019).

The association between learners' forum participation and dropout in MOOC is a complex issue which has been analyzed through a quantitative dimension (e.g., Chen & Zhang, 2017; Pursel et al, 2016), a content dimension (e.g., Capuano et al., 2021; Galikyan et al., 2021), and a temporal dimension (e.g., Tang et al., 2020; Yang et al., 2022). In this paper, the quantitative dimension refers to number of the posts; the content dimension refers to contents of the posts; and the temporal dimension refers to the continuity of posting behavior and time sequence. Concerning the quantitative dimension, learners with more posting activity are less likely to drop (Pursel et al, 2016). Furthermore, learners' super-posting behavior can lead to better adamic achievement (Huang et al., 2014). From the content dimension, the act of posting course-related posts continuously can lead to high retention rate (Yang et al., 2022). The non-completers tend to express course-unrelated matters in course reviews (Peng & Xu, 2020). From the temporal dimension, learners' persistent forum participation can lead to better performance including higher retention rate and grades (Tang et al., 2018; Yang et al., 2022). Most of the previous research applied forum participation as a simple feature to predict dropout through a single dimension, which is insufficient (Moreno-Marcos et al., 2020). Better understanding of features is required before applying them in prediction models (Chen et al., 2019).

Machine learning has been widely applied in MOOCs with a specific focus on predicting learner dropout in MOOCs (e.g., Balakrishnan & Coetzee, 2013; Chen & Zhang, 2017). In particular, the hidden Markov model (HMM) has proved to be an effective way to predict dropouts when applied to MOOCs (e.g., Balakrishnan & Coetzee, 2013; Fei &

Yeung, 2015; Geigle & Zhai, 2017). An HMM is a probabilistic generative model to determine the best sequence of model states based on the inputs of a sequence of observations (Rabiner, 1989). Thus, it is a suitable method to model sequence data and understand the latent behavior patterns behind the data. Balakrishnan & Coetzee (2013) first applied HMM to understand learners' behavior and predict dropouts. Applying HMM in MOOCs allows a speculation on learners' behavior in the next time step based on their current actions and previous states (Balakrishnan & Coetzee, 2013), which enables us to analyze the MOOC learners' forum behavior through multiple dimensions.

Although previous studies have explored learners' forum activities (e.g., Chen & Zhang, 2017; Galikyan et al., 2021; Yang et al., 2022), what is not yet clear is how different forum behaviors at different stages associate with learner dropout. Thus, we decided to analyze the forum behavior through a multidimensional analysis, which includes a quantitative dimension, a content dimension and a temporal dimension. This paper has two aims. Firstly, it explores the association between learners' forum behaviors and dropouts. Secondly, it aims to describe the design and implementation of HMM on predicting dropouts based on forum participation data in MOOCs. The research questions this paper aims to answer are as follows:

RQ1: How do MOOC learners' forum participation behaviors associate with their dropout?

RQ2: What are the hidden states behind MOOC learners' forum participation behaviors? How does each state associate with dropouts?

RQ3: To what extent can learners' forum participation HMM states improve the predictive power of current MOOC dropout prediction models?

4.2 Literature Review

4.2.1 MOOC Forum participation and dropout

It is now well established that forum participation is positively correlated with learners' course completion, and it can lower learner dropout rate (Chen & Zhang, 2017; Diver & Martinez, 2015; Semenova, 2022). Learners were even more likely to give up if they did not participate in the discussion forum (Tang et al., 2020). From a quantitative perspective, the total number of posts learners posted in the forum influenced their probability of completing the course. For example, Pursel et al.'s study (2016) revealed that each additional post per week increased the rate of completion by 3.1%. Although forum participation is not a measurement of course grade, forum posts and comments are "significant predictors of completion" (p. 213). Furthermore, the content of the forum posts can be another dimension to influence learners' course completion. The large number of participants in MOOCs causes an information overload and chaos in MOOC forums (Brinton et al., 2014). The chaos may cause a lack of necessary support which can lead to dropouts (Alrajhi et al., 2020; Capuano et al., 2021). To avoid the chaos and further explore the contents in forums, previous studies have primarily concentrated on identifying the contents of the posts (Alrajhi et al., 2020; Brinton et al., 2014; Capuano et al., 2021; Wise et al., 2017; Pillutla et al., 2020). In a recent study, Gamage et al. (2020) founded that 79% of the forum participation was related to technical matters or assignment related matters. Tang et al.'s (2020) study further showed that learners were more likely to drop the course if they only used the forum to make social communication or post course management related questions. From a temporal perspective, persistent forum participation can lead to better performance (Yang et al., 2022). While temporal analysis has already been widely applied to dropout prediction (e.g., Boroujeni & Dillenbourg, 2018; Moreno-Marcos et al., 2020), few of the previous studies analyzed forum participation through a temporal dimension. It remains unclear what kind of learner behavior is associated with dropout. This topic requires further exploration. It is not enough to consider the forum participation as a feature for dropout prediction only from a single dimension (a quantitative dimension or a content dimension). It is also important to understand the learners' behavior during the course so that course facilitators can pinpoint the best moment to take corrective actions.

4.2.2 MOOC dropout prediction

In the past decades, researchers have been exploring innovative ways to promote MOOCs learners' engagement and improve their learning experience (e.g., Greene et al., 2015; Goel & Goyal, 2020). Greene et al. (2015) stated that attrition in MOOCs happens over time. As time passed, the proportion of learners who dropped decreased. Firmin et al. (2014) proposed that the online support services provided by the faulty had a significant positive association with learners' course completion. However, this kind of intervention may have high variable costs and be very time intensive. Moreover, it is difficult to know the reasons for dropping out since the learners who dropped were unlikely to take the post-survey (Eriksson et al., 2017).

Thus, researchers seek machine learning as a new way to explore the dropout issues. Previous research suggests using machine learning methods for dropout prediction based on learners' demographics, clickstream and events (e.g., Chen & Zhang, 2017; Goel & Goyal, 2020; Whitehill et al., 2015; Moreno-Marcos et al., 2020). In an early study, Whitehill et al. (2015) built a dropout classifier using multinomial logistic regression with five features, including persistent time, number of different events (e.g., forum posts, video plays), grades and so on. They further developed a dynamic survey intervention system. It is worth noting that their interventions induced learners to come back sooner into the course. Moreno-Marcos et al. (2020) applied four classical machine learning algorithms with seven different features to explore the predictive power of self-regulated learning (SRL) strategies in dropout prediction. They found the models built based on event based SRL strategies achieved good performances. They also reported that the best moment to predict dropout is the second week of the MOOC. In addition, clickstream and demographics have been combined to predict dropouts (Goel & Goyal, 2020). In their study, they reported that clickstream contributed more to dropout prediction. To include the temporal dimension, temporal models including HMM and recurrent neural network (RNN) model have been proposed to predict dropout in MOOC (Fei & Yeung, 2015). Their results showed that RNN models with long short-term memory cells obtained significantly better dropout prediction results than HMMs. In contrast, Chen et al. (2019) argued that deep learning methods were better than the traditional machine learning methods on dropout prediction. They noted that deep neural networks required iterative training and a large amount of training data, which may not be available in a single MOOC. Also, the lack of understanding of learning behaviors in MOOCs may lead to un-unified conclusions on behavior features, which may lead to better but biased classification results (Chen et al., 2019). Thus, it is important to understand the learners' forum participation through multiple dimensions and examine its power as a prediction feature for dropout.

4.3 Methodology

4.3.1 Dataset

The data used to build the HMMs was extracted from the discussion forum of the Introduction to Art: Concepts & Techniques. It was a seven-week MOOC with up to 69,867 learners enrolled. In total, 7,292 learners participated in the forum during the seven weeks period. Among the learners who participated in the forum, 37% (2,723) of them did not complete the course. In total, 29,111 posts and 22,040 comments were extracted from the forum. We did not distinguish posts and comments in this research. All the term "posts" imply both posts and comments in the following sections.

4.3.2 Content-related / Content-unrelated posts

Before building the models, we conducted a content analysis on the 51,251 posts. We adopted the same procedure for building classification models as that presented by Yang et al. (2022). First, we applied latent semantic analysis on the contents of the forum posts. Then, we used a decision tree model to classify the posts into content-related posts and content-unrelated posts. The content-related posts include posts related to course topics (e.g., self-introductions, discussions about artwork, questions and comments about the course contents, and assignment questions). In this course, learners are required to talk about their personal experience with art in the self-introduction. Thus, self-introductions are included as a content-related topic. Content-unrelated posts are posts that contained technical problems, course administration issues, course evaluations, complaints and so on. The training set for the classification model was composed of 4,108 posts which were chosen randomly and coded by two researchers. The two researchers reached an agreement statistic of .82 (Kappa value), which is an acceptable level of interrater reliability. The classification model reached an accuracy of .76 with an AUC value of .83 and a Kappa value of .506. After classification, 38,340 posts were labeled as content-related posts and 12,911 posts were labeled as content-unrelated posts. The full procedure to build the classification model are introduced in the paper of Yang et al. (2022). The examples of content-related and content-unrelated posts are shown in Table 4.1.

4.3.3 Dropout definition

We followed the definition of dropout of Yang et al. (2022), which is an integrated dropout definition of Nagrecha et al. (2017) and Moreno-Marcos et al. (2020). In this course, a learner needs to gain 80 out of 100 points to get a certification. Thus, a learner is considered dropped if he/she: 1) stopped accessing the course before the final week of the course; and 2) received less than 56 points in total.

4.3.4 Time unit and time period

Previous research using temporal analysis on MOOCs was mainly based on a weekly basis (e.g., Chen & Zhang, 2017; Moreno-Marcos et al., 2020). For the specific MOOC we analyzed, lectures and activities were given weekly and most of the learners posted less than eight posts per week. Thus, we decided to use each week as the time unit.

The dropout rate varied over time (Greene et al., 2015). The learners’ forum behavior changes as the course progresses. We observed that there were different dropout rates, finish rates and behavior patterns in early and late phases of the course, which might lead to different transition probabilities. In our process of building a seven-week model, we noticed that this model cannot reflect the precise probabilities of the transitions between different latent states as these change from early in the course to later in the course. In addition, one latent state might be interpreted differently at different phases of the course. To get more specific probabilities of the transitions in different periods of the course, we further separated the course into two periods (early phase and late phase) according to learners' activity patterns. We then built an HMM for each separate period. In total, we built three HMMs representing the three periods of the course. The first model was built based on the forum data of the whole course (seven weeks). The second model was built in the early phase of the course (first four weeks). The third model was built based on the late phase of the course (last three weeks) of the course.

4.3.5 Preprocessing of observations

HMMs require observation sequences as the inputs. Thus, we mapped the number of each learner's posting activities in each week to the observations in the sequence data. To categorize the number of observations, we first developed the content-related/content-unrelated involvement levels according to the activities' distribution (see Table 4.2). We then mapped different numbers of content-related/content-unrelated posts to the involvement levels. After that, to investigate both course-related and course-unrelated participations, each learner's content-related and content-unrelated involvement levels were combined as the learner's observation in that week. For example, if a learner posted four content-related posts and one content-unrelated post, then his/her posting observation in that week is medium-medium. For dropouts, we added a drop observation after the observation in the last active week. Similar to the drop observation, if a learner achieved more than 80% of the final grade or reached the final week of the course, a finish observation was added after that week's observation. For the early phase model, if the learner did not drop the course, a retention observation will be added after the fourth observation. In total, there were 23 different observations in our models.

4.3.6 HMM model generation

We built our HMMs based on the open-source library “Jahmm” (Francois 2006). Jahmm provides a Viterbi algorithm function to decide the most likely sequence of hidden states corresponding to a given observation sequence. For model selection, we first applied Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as the measurements. However, the results showed that both AIC and BIC favored the models which had the minimum number of states (two). However, the two-state model cannot separate the dropout behaviors and only two states were insufficient in explaining the latent behavior patterns behind forum participation. Thus, we moved our model selection design from AIC, BIC to a traditional evaluation procedure, i.e., k-fold cross-validation.

4.3.7 Design of the baseline models and improved models

We considered five features to build the baseline dropout prediction models. The set of features include five high predictive power event-based features extracted from learners’ clickstream data (Kloft et al., 2014; Moreno-Marcos et al., 2020): 1) total server requests; 2) total sessions; 3) total page views 4) total videos watched 5) total video actions. Each learner’s weekly clickstream events are extracted and converted into the features listed before. We applied Support Vector Machines as the predictive algorithm since it is a classical machine learning algorithm widely applied in previous dropout prediction tasks (e.g., Jin, 2020; Kloft et al., 2014; Moreno-Marcos et al., 2020). In addition, another two prediction models were built based on random forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016) to further test the prediction power of the clickstream events. The models were built based on the algorithm packages in Weka (Witten et. al., 2016). For the datasets used to train the models, we only included the clickstream data produced by the learners who participated in the forum. To avoid having imbalanced datasets, the ratios of retention and drop for the two datasets were set to 2:1 and 1:1. In total, there are 5,440 records in the 1:1 dataset and 8,160 records in the 2:1 dataset. The models were tested based on the 5-cross validation.

To test how learners' forum participation behavior can improve the accuracy of the dropout prediction tasks, we further applied the three machine learning algorithms to build six prediction models with 7 weeks forum participation states as additional features. In addition, to demonstrate the importance of the continuous forum participation states, we also included the learners' forum states in the previous week and built another six improved prediction models.

4.4 Results

4.4.1 Seven-week model

We applied a 10-fold cross-validation process and generated nine preliminary HMMs using the seven weeks of forum data in order to explore HMMs ranging from two states to ten states. As can be seen in Table 4.3, from cross-validation result, the nine-state HMM was the first one that maximized the average likelihood and had a reasonable number of states in describing learners' forum behaviors. Thus, we decided to use the nine-state HMM as the seven-week model. The latent state transition diagrams are shown in Figure 4.1.

4.4.1.1 Interpretations of the latent states in the seven-week model

We then interpreted the nine latent states in the seven-week model according to 1) the observations captured by each state and 2) the transition probabilities between states. In particular, we pinpointed each state's transition probabilities to the dropout state and the finish state in Figure 4.7(a). The weekly frequency of each state is shown in Figure 4.2(a) and their weekly proportions are shown in Figure 4.2(b). Additionally, for each state, we identified the number of students that will transition to the drop state in the following week. This drop fraction is defined as (# of students in a state that will drop in the following week)/(# of learners currently in that state). For each state we also calculate its fraction of the total number of students in all states that will drop in the following week, i.e., as (# of students in a state that will drop in the following week)/(total # of dropouts in the following week) for each state are shown in Table 4.4 and Figure 4.3.

State IA. State IA refers to "Inactive State". The only observation associated with state IA involved no content-related posts and no content-unrelated posts, which indicated that all the learners in state IA did not participate in the forum that week. The total number of learners in state IA increased significantly after the first week and decreased slightly in the last week. Except for the first week, around 80% of the learners stayed in state IA in the early phase of the course. More than 90% of the learners stayed in state IA in the late phase of the course. For state IA, the transition probability to finish state was .1 and the transition probability to drop state was .05. Besides, after the first two weeks, almost all the dropouts (over 90%) were in state IA.

State A. State A refers to "Active State". All the observations captured in this state were associated with a medium level of content-related involvement. Half of the learners in state A did not post unrelated posts that week. Others posted at least one unrelated post in that week. The total number of learners in state A decreased as the course went on. For state A, the transition probability to drop state is .06 and the transition probability to finish state is .03 which was below the average.

State F. State F is the "Finnish State". It captures all the observations of students completing the course.

State Pas. State Pas was described as "Passive State". In state Pas, 99.6% of the observations involved posting one or two content-related posts and no content-unrelated posts, indicating that learners in this state posted less than two posts in that week. As shown in Figure 4.2(b), State Pas accounted for a significant proportion in the early phase of the course compared to the other states (except for state IA). In addition, as shown in Figure 4.1, the initial probability for state Pas is .38, which showed that almost 40% of the learners joined the forum through this state. The total number of learners in state Pas kept decreasing in the early phase of the course and remained at a relatively stable level at the late phase of the course. The transition probability of dropping state is .09, which is the highest one among the states. Besides, as shown in Figure 4.3 and Table 4.4, at the first week of the course, 11% of the learners in Pas State dropped, which occupied almost 60% of the dropouts in that week. Although the total number of dropouts in state Pas decreased in the following weeks, its proportion remained high compared to the other states. The transition probability to finish state is .05, which means learners were unlikely to finish the course when they were in this state.

State Prob&E. State Prob&E (Problem-Evaluation State) captured the learners who posted only unrelated posts. Given the different contents of the unrelated posts (e.g., problems, course evaluations), state Prob&E might represent different behaviors in the early phase of the course and the late phase of the course. Therefore, it refers to "Problem-Evaluation State". The total number of students in the state Prob&E kept decreasing during the first six weeks, but then increased hugely during the last week. The transition probability to drop state is .06 and the transition probability to finish state is .18, which indicates that in the late phase of the course, the learners in state Prob&E were more likely to finish the course.

State D&S. State D&S refers to "Drop and Super-poster State" as 96% of the learners in this state had a drop observation. The rest 4% of the learners were super-posters. From Figure 4.1 and Figure 4.2(b), we can see that about 1% of the learners were super-posters and most of them had super-posting behavior in the early phase of the course. We noticed that all the transition probabilities related to the super-posters were less than .01 transfer from other states to superposter, thus the model did not return the specific probabilities. However, we speculate that the super-posters would transfer to another state after posting a huge number of posts in one week.

State Pas& Prob. State Pas&Prob was similar to state Pas and was named "Passive with Problems State". The only difference between state Pas&Prob and state Pas was that the learners in state Pas&Prob posted one more unrelated post in that week. However, compared to state Pas, the transition probability to drop state decreased to .05 and the transition probability to finish state increased to .1. In addition, the dropout rate of state Pas&Prob is lower than both state Pas and state Prob&E.

State Pas&E. State Pas&E refers to “Passive and Evaluation State”. Few learners were in this state during the course. All the learners in state Pas&E posted less than two related posts. In contrast, they posted more than two unrelated posts. For state Pas&E, the transition probability to drop state is .03 and the transition probability to finish state is .11. Compared with state Pas&Prob, learners’ behavior in state Pas&E lead to a lower drop rate and a higher complete rate, which is more likely to be the course evaluation behavior. Thus, we mentioned state Pas&E as “Passive and Evaluation”.

State Pos. State Pos refers to “Positive State”. Other than super-posters in state D&S, the learners who posted more than six posts were captured in state Pos. All the learners in state Pos posted more than 10 posts in that week, which was highly above average. The transition probability to drop state is .04 and the transition probability to finish state is .03, which indicates that this posting behavior is unlikely to appear before they finish. Similar to the super-posters, learners in state Pos had an intense posting behavior in the early phase of the course.

4.4.1.2 States patterns before drop

According to our findings, the associations between drop state and other states cannot perfectly reflect the learners’ behaviors before they drop. Hence, we suggested several observable latent states patterns which can better reflect dropouts’ behaviors. The learners’ states patterns before dropout at each week were counted. The proportion of the two sequences patterns (those ever greater than 2%) and three sequences patterns (those ever greater than 3%) before dropout at each week are shown in Figure 4.4(a) and Figure 4.4(b). As can be seen, the patterns associated with state Pas and state Prob&E always kept a significate proportion among all the patterns. This result showed that some dropouts after state IA might be leaded by state Pas or state Prob&E.

4.4.2 Early phase model

The early phase HMM was built upon the first four weeks' forum data. We noted that a new observation "retention" had been added since the learners' finish state might not be decided at the fourth week. The 10-fold cross-validation result showed that the eight-state model was the first one to reach the maximum average likelihood with a reasonable number of states in describing learners' forum behaviors. The latent state transition diagrams for the eight-state early phase model are shown in Figure 4.5. The transitions to the drop state and finish or retention state are shown in Figure 4.7(b).

4.4.2.1 Interpretations of the latent states in the early phase model

State I. The state I in the early phase model is the same state as in the seven-week model. The transition probability to drop state increased to .08 and the transition probability to retention state is .24.

State A. Compared to the seven-week model, the new attribute included in the state A was the "low-super" observations. However, the proportion of the "low-super" observations in state A is less than 2%. Thus, we regarded state A in the early phase model as the same latent state as the state A in the seven-week model.

State R. A portion of 98.9% of the learners in state R had the observation of "retention" which means they kept learning after the fourth week. A number of 51 learners finished the course before week four and they were also captured by this state.

State Pas. The state Pas in the early phase model is the same state as the state Pas in the seven-week model. The transition probability to drop state slightly increased to .1 and the transition probability to state R is .7.

State Prob. The state Prob in the early phase model has the same elements as the state Prob&E in the seven-week model. The only difference is that the learners were not required to evaluate the course in the early phase of the course. Thus, the unrelated posts were unlikely to associate with course evaluation. We concluded that the learners in this state did not post content-related posts. Instead, they posted questions or unrelated information. For the new state Prob in the early phase model, the transition probability to drop state increased to .8 and the transition probability to retention state hugely decreased to .13, which validated our conclusion.

State D&S. The state D&S in the early phase model is the same state as the state D&S in the seven-week model. Thus, it still represents the drop behavior and super-posting behavior.

State Pas&Prob. The new attribute for the state Pas&Prob is the “low-high” observation (posting one or two content-related posts and two or three content-unrelated posts), which does not change the interpretation of the state Pas&Prob. The transition probability to drop state slightly increased to .6 and the transition probability to retention state hugely decreased to .14. In the early phase model, the difference between the state Pas&Prob and the state Prob is that learners still posted related posts in the state Pas&Prob which led to a lower dropout rate and a higher retention rate.

State Pos. The state Pos in the early phase model is the same state as the state Pos in the seven-week model. The transition probability to drop state remains the same (.04) and the transition probability to retention state hugely increased to .09 (used to be .03). This result indicates that the positive behaviors were more likely to occur in the early phase of the course.

4.4.3 Late phase model

The last HMM was built based on the last three weeks' forum data. The 10-fold cross-validation result showed that the 5 states HMM was the first one to reach the peak of average likelihood (.063). The latent state transition diagrams for late phase model are shown in Figure 4.6. The transitions to finish or drop state are shown in Figure 4.7(c).

4.4.3.1 Interpretations of the latent states in the late phase model

State I. The state I in the late phase model is the same state as in the previous models. As shown in Figure 4.6, the initial probability for the state I is .87 which indicates that most of the learners remained inactive at this phase of the course. For state I, the transition probability to finish or drop state is .34.

State F&D. State F&D in the late phase model represents the finish or drop activities. 89% of the learners in this state finished the course and 11% of the learners dropped. Judging by the HMM result, the drop activity and finish activity had no difference in the late phase of the course.

State E. State E in the late phase model refers to "Evaluation State". All the learners in this state posted only unrelated posts which were likely to be course evaluations. For state E, the transition probability to finish or drop state is .55, which might be because of the learners.

State A&E. State A&E refers to "Active and Evaluation State". Most of the learners in state A&E remained active at the end of the course and posted considerable related posts and unrelated posts (most likely course evaluations). The transition probability to finish or drop state is .35.

State Pas&E. The state Pas&E in the late phase model is the same state as in the seven-week models. In the late phase model, the difference between state Pas&E and state E was that learners in state Pas&E still posted one or two related posts. It might be because they were actively participating in the forum all the time, or because their course evaluation posts contained content-related information. Thus, we named this state "Passive and Evaluation". The transition probability to finish or drop state for state Pas&E is .43.

4.4.4 Dropout prediction models

The evaluation metrics for the dropout prediction models include accuracy, F1, and AUC value. The results of the baseline models and improved models are shown in Table 4.5. We can see that the performance of the clickstream-based dropout prediction models can be improved with the feature of learners' forum participation states. Especially, with the help of the learners' two weeks' forum participation states (the current week and the previous week), the accuracy of the dropout prediction models can be increased by around .02. In addition, among the models built with different machine learning algorithms, the prediction model built with XGBoost achieved the best evaluation result with an accuracy of .729.

4.5 Discussion

This paper aimed to analyze the association between the MOOC learners' forum participation behaviors and their dropout behaviors. For that, three hidden Markov models were developed according to different course periods. Based on the results of the three models, we discuss the associations between learners' forum participation behaviors and their dropout combines three dimensions: a quantitative dimension, a content dimension, and a temporal dimension.

The model results show that a sparse posting behavior (posting only one post in a week) and a zero-posting behavior are greater signs of dropout. In addition, the behavior patterns show that there is a path from sparse posting behavior to zero-posting behavior and finally a dropout. These results show that learners were more likely to give up once they lost their interest in the discussion forum or stopped participating in the discussion forum. In another study, Semenova (2022) proposed that the participants who took the course with the purpose of getting access to the course materials are less likely to complete the course, which is also a reasonable explanation for these kinds of dropping behaviors.

Another contribution of this paper is that we explored the super-posting behavior. As can be seen from the model results, the state of the super-posting behavior and the drop behavior cannot be separated. We speculated that it is because there were almost no transitions from the super-posting behavior to the drop behavior. As can be seen from Table 4, only one learner dropped the course after a super-posting behavior. This result shows that the super-posting behavior can significantly reduce the probability of dropping in the coming week. As Huang et al. (2014) stated, super-posters commonly have better learning performance. Our results extend their results by showing that super-posting behavior can also reduce the learner's drop rate. In addition, as shown from the result of the transition probability from state D&S to state A, there also exists a posting decline in the super-posting behavior. From a temporal dimension, super-posters may transfer to normal posters. It is very important to analyze the behavior from a multidimensional analysis.

From the perspective of dropout prediction, we explored the association between unrelated posts and learners dropping out. The early phase model shows that the learners were more likely to drop the course if they only used the forum to post course management related questions, which coincides the previous findings of Tang et al. (2020). However, it is remarkable that the learners who posted both content-related and content-unrelated posts have a higher retention rate than those who only posted content-related or content-unrelated posts. The content-related posting behavior together with some course management related posts might show the learners' intention to keep learning in the course.

Across the three models, several states were found to be significant. The state Pas (Passive State) and state Prob (Problem State) had higher transaction probabilities to state D&S (Drop State). Previous research suggests that the total number of posts can be used to predict dropout (e.g., Balakrishnan & Coetzee, 2013; Chen & Zhang, 2017; Whitehill et al., 2015; Pursel et al, 2016;), which may lead to a misconception that the learners who posted one post in that week (passive) are less likely to drop compares to those who posted zero posts (inactive). However, in our results, compared with the state Pas, the state IA (Inactive State) becomes a less prominent indicator of dropout. This result indicates that after participating in the MOOC forum, the starting point of a dropout is the sparse posting behavior, not the zero-posting behavior. It may be because the learners' drop decision has been made before their non-participating behavior. We agree with Chen et al. (2019) that it requires more understanding of learners' behavior before applying them as a feature for prediction. Even considering one post for each week at the early phase of the course (which means consistent participation), students may not complete the course. The learners' forum behavior should be analyzed through multiple dimensions instead of a single dimension which may lead to one-sided conclusion. In addition, compared to state Pas and state Prob, a lower drop rate of state Pas&Prob indicates that learners who posted both content-related and content-unrelated posts are less likely to drop the course. This result indicates that the content-unrelated posts could also be a feature for dropout prediction. It is remarkable that the late phase model did not separate the drop behavior and the finish behavior into two states, which indicates that little drop behaviors happened in the late phase of the course. Thus, the best time to provide intervention is in the early phase of the course (first four weeks for a seven-week MOOC). From the perspectives of both the transition probabilities and the contents of the posts, it is clear that the best state for the learners to stay in is state A (Active State) and State Pos (Positive State).

The drop rate differed by time period (Greene et al., 2015). We created three HMMs for the same MOOC with forum data in three different time periods. The different results of states and transition probabilities in the three models show that the learners' posting behavior and drop rate varied at different periods of the course. Although the seven-week model can provide a holistic picture of the learners' behavior patterns over the course, it may not reflect the learners' behaviors at a specific time properly. Creating two more models enabled us to rethink learners posting behavior and drop rate at different phases of the course. This result further supports the idea of Chen and Zhang (2017) that future dropout prediction model should be built upon a short-term period. When considering the HMMs as a plugin for the MOOC platforms, it is better to build the model based on a short-term period of data. The advantage of the early phase model is that it can clearly locate the learners with problems (state Prob). From our data, the late phase model cannot separate the finish behavior and the drop behavior. Thus, from the perspective of dropout prediction, the late phase model is not necessary. Similar to the findings of Moreno-Marcos et al. (2020), the hidden Markov model should be built based on the forum data of the early phase of the course in order to predict the dropout and find the learners who seek help.

To test the prediction power of the forum participation states, we further developed six models for dropout prediction. The model results show that forum participation is an important feature to predict dropout. All the improved dropout prediction models achieved better evaluation results compared to the baseline models. While a single state cannot tell the whole picture, the continuous forum participation states can significantly increase the accuracy of the dropout prediction models.

4.6 Practical implication and future works

This study has implications for MOOC instructors, facilitators and researchers. Through a multidimensional analysis, our study suggests that instructors should rethink learners' forum behavior and dropout from a quantitative dimension, a content dimension, and a temporal dimension. First, instructors should pay more attention to those learners with sparse posting behavior and zero-posting behavior which may lead to a drop. After analyzing the models' results, we found that the best time to prevent drops is the early phase of the course. MOOC facilitators need to start intervention at the beginning of the course. The interventions may include creating more forum activities and tasks to encourage forum participation, requiring peer-to-peer review and discussions, promoting instructor-learner interactions, providing regular email notifications and conversational agents (Capuano et al., 2021; Galikyan et al., 2021; Gamage et al., 2020; Pillutla et al., 2020). As shown by our results, a course evaluation behavior in the late phase of the course also reduced the drop rate, which is consistent with Peng and Xu's (2020) result that posting reviews is a significant behavior for MOOC completers. Asking the learners to provide their feedback after every lecture could be another possible intervention to promote forum participation. Second, the short-term HMMs can be applied to the current MOOC platforms as it can help identify the states of the learners. Instructors and facilitators can use it as plugin to predict dropouts. After applying the model, the course facilitators can easily locate the learners who need additional help through the states and provide assistance accordingly.

This study suggests several implications for researchers. First, when including forum participation as a feature, researchers should consider it through multiple dimensions to get a comprehensive conclusion. Second, the forum posts described in this paper were only classified into two categories. A more precise classification of posts may lead to more behavior patterns and more accurate models. Thus, future research can improve the model by applying a more precise classification of posts. For example, providing more specific categories of contents or classifying the posts based on Chi's ICAP (Interactive-Constructive-Active-Passive) framework (Chi & Wylie, 2014). In addition, this study provides a pipeline of building HMMs, which can be trained based on the forum data for a single MOOC. Like previous research (e.g., Chen & Zhang, 2017; Geigle & Zhai, 2017), with accessible forum data, the training process can be generally applied to other courses in other domains. Future work can apply the model into other domains to test the generality and transferability of the model.

4.7 Conclusion

This study explored the association between forum behavior and dropout from multiple dimensions through HMM. It provides a pipeline of building an HMM which can be applied to the MOOC platforms to better support the facilitators. The results of this paper showed that forum behavior can be applied to predict dropouts and identify the learners' status. Especially, the sparse posting behavior was identified as a sign of dropout. Another major finding is that learners' forum behavior differs by course periods, which requires short-term prediction models. These results expand our understandings of learners' forum behavior and dropout and urge us to rethink the learners' forum behavior as a feature to predict their drop from multiple dimensions. The total number of posts and the contents of posts need to be combined with the temporal dimensions in order to get a comprehensive picture of learners' retention. It is hoped that this research will help researchers and instructors to rethink the importance of forum behavior and its power to predict dropouts.

Table 4.1 Examples of content-related and content-unrelated posts

Content-related examples	Category	Content-unrelated examples	Category
I start are since my childhood, my father is art enthusiastic, my sister too practice, and I was one of best vehicle painter during my schooling period.	Self-introductions	The problem summary: When I wanted to look at Student 2 work, it always shows me server error. Can it be fixed?	Technical problems
I am posting a watercolor that I painted about 4-5 months ago. I used the pen and wash method. It was probably the first painting that turned out well for me. I am passionate about watercolors and am learning to use new media as well. This course is very exciting for me!	Discussions about artwork	I really like this course, but the artist videos are very disappointing. They could show many more examples and would be much better if read by someone concentrating more on the subject than their own diction. On the other hand, the demonstration videos are terrific.	Course Appraisal
This is so imaginative, wonderful work! Is the hair made from the filaments?	Discussions about artwork	I am having the same problem and it makes me sad I spent so much time doing my artwork. I really would like this to get fixed.	Course administration issues

Table 4.2 Involvement levels

Content-Related		Content-Unrelated	
Number of Posts	Involvement Level	Number of Posts	Involvement Level
0	none	0	none
1-2	low	1	medium
3-7	medium	2-3	high
8-16	high	more than 3	super
more than 16	super		

Table 4.3 Seven weeks 10-fold cross-validation result

	2 States HMM	3 States HMM	4 States HMM	5 States HMM	6 States HMM	7 States HMM	8 States HMM	9 States HMM	10 States HMM
Average Likelihood	0.0003	0.0013	0.0014	0.0016	0.0017	0.0018	0.0018	0.0019	0.0019
Lowest Likelihood	0.0002	0.0012	0.0013	0.0014	0.0015	0.0016	0.0016	0.0017	0.0017
Highest Likelihood	0.0003	0.0014	0.0015	0.0017	0.0018	0.002	0.002	0.002	0.002
Standard Deviation	2.73×10^{-4}	5.98×10^{-5}	6.46×10^{-5}	7.24×10^{-5}	8.17×10^{-5}	1.16×10^{-4}	9.37×10^{-5}	9.49×10^{-5}	9.54×10^{-5}

Table 4.4 Weekly transitions from each state to drop state with drop fractions

		Early Phase			Late Phase	
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
IA	10 (0.65%/2.18%)	431 (9.15%/74.05%)	489 (9.09%/90.06%)	335 (5.8%/92.03%)	267 (4.46%/95.02%)	216 (3.5%/97.30%)
A	83 (7.17%/18.08%)	21 (5.21%/3.61%)	9 (3.98%/1.66%)	4 (2.96%/1.10%)	1 (1.32%/0.36%)	2 (3.92%/0.90%)
Pas	271 (10.96%/59.04%)	93 (11.06%/15.98%)	24 (5.94%/4.42%)	9 (3.25%/2.47%)	5 (2.76%/1.78%)	1 (0.79%/0.45%)
Prob&E	48 (10%/10.46%)	17 (7.46%/2.92%)	12 (4.86%/2.21%)	10 (6.58%2.75%)	6 (3.7%/2.14%)	1 (1.1%/0.45%)
D&S	1 (1.45%/0.22%)	0 (0%/0%)	0 (0%/0%)	0 (0%/0%)	0 (0%/0%)	0 (0%/0%)
Pas&Prob	30 (6.98%/6.54%)	15 (7.98%/2.58%)	5 (3.65%/0.92%)	5 (4.5%/1.37%)	2 (3.03%/0.71%)	1 (1.67%/0.45%)
Pas&E	7 (6.6%/1.53%)	3 (4.11%/0.52%)	3 (4.17%/0.55%)	0 (0%/0%)	0 (0%/0%)	1 (2.5%/0.45%)
Pos	9 (1.96%)	2 (0.34%)	1 (0.18%)	1 (0.27%)	0 (0.00%)	0 (0.00%)

Table 4.5 Dropout prediction models evaluation results

Dataset 5 cross-validation	Model	Forum states included	Accuracy	F1	AUC
1:1 (5440 records)	SVM	No	0.626	0.624	0.626
		Yes (single states)	0.628	0.626	0.628
		Yes (two states)	**0.643**	**0.642**	**0.643**
	RF	No	0.637	0.636	0.691
		Yes (single states)	0.652	0.651	0.716
		Yes (two states)	**0.663**	**0.662**	**0.724**
	XGBoost	No	0.652	0.65	0.714
		Yes (single states)	0.665	0.662	0.737
		Yes (two states)	**0.678**	**0.677**	**0.745**
2:1 (8160 records)	SVM	No	0.701	0.669	0.603
		Yes (single states)	0.71	0.673	0.606
		Yes (two states)	**0.721**	**0.691**	**0.624**
	RF	No	0.689	0.674	0.677
		Yes (single states)	0.699	0.683	0.702
		Yes (two states)	**0.715**	**0.698**	**0.729**
	XGBoost	No	0.707	0.678	0.712
		Yes (single states)	0.711	0.683	0.729
		Yes (two states)	**0.729**	**0.706**	**0.751**

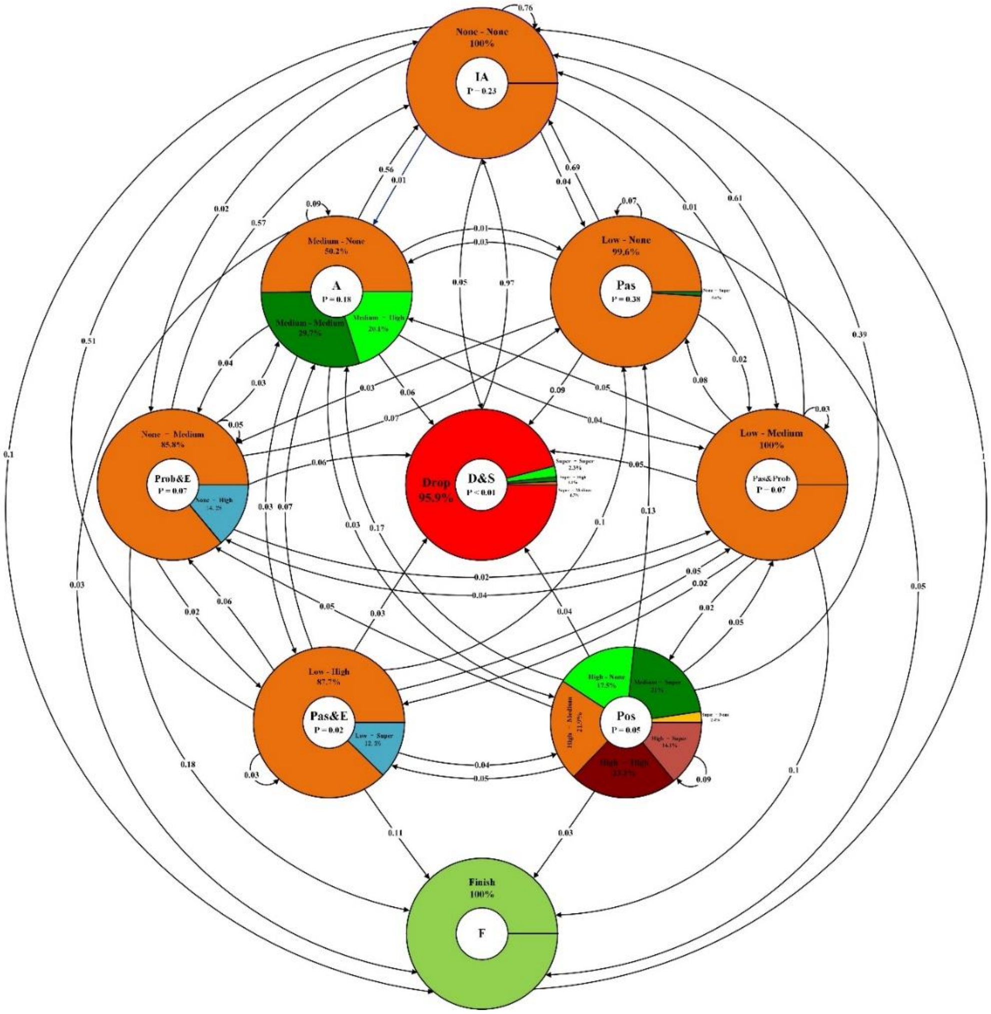


Figure 4.1 The latent state transition diagrams for seven-week model (transitions over 0.01)

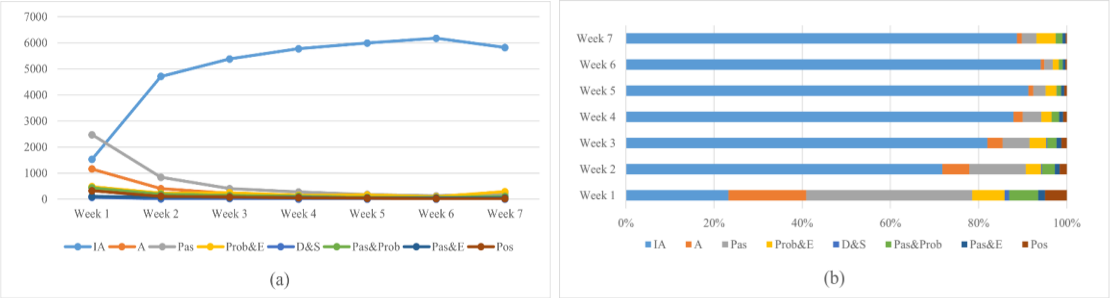


Figure 4.2. (a) Weekly frequency of each state (b) Weekly proportion of each state

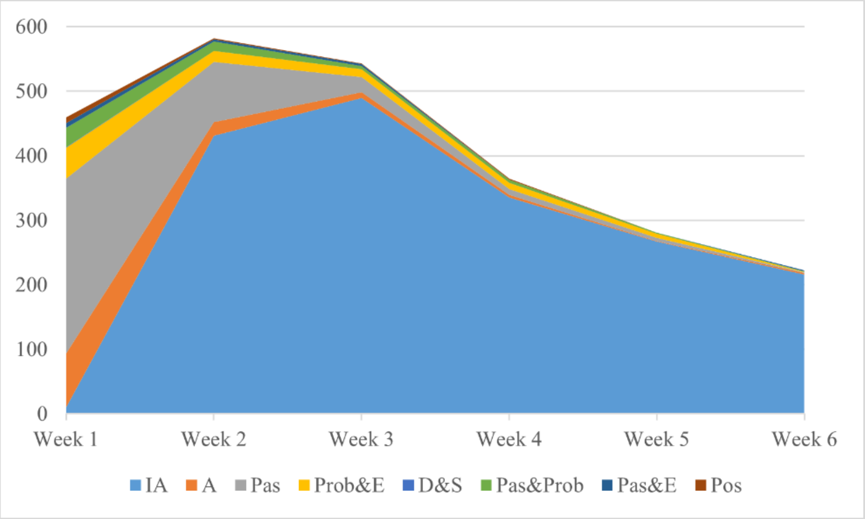


Figure 4.3 Total number of states before dropout

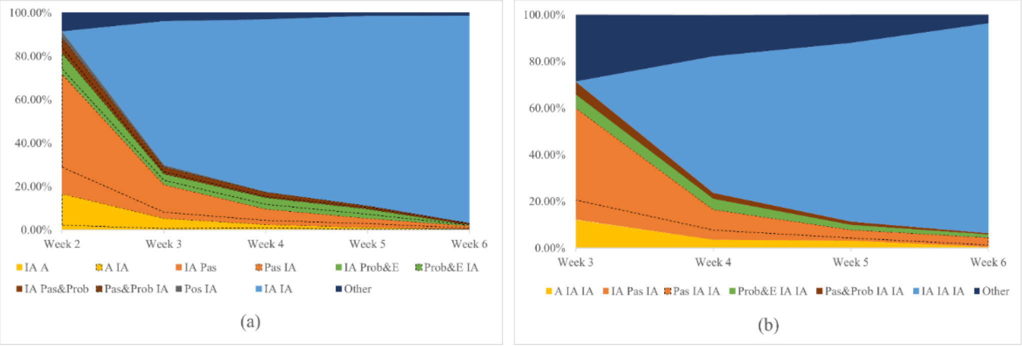


Figure 4.4 (a) Proportion of two sequences pattern and (b) Three sequences pattern

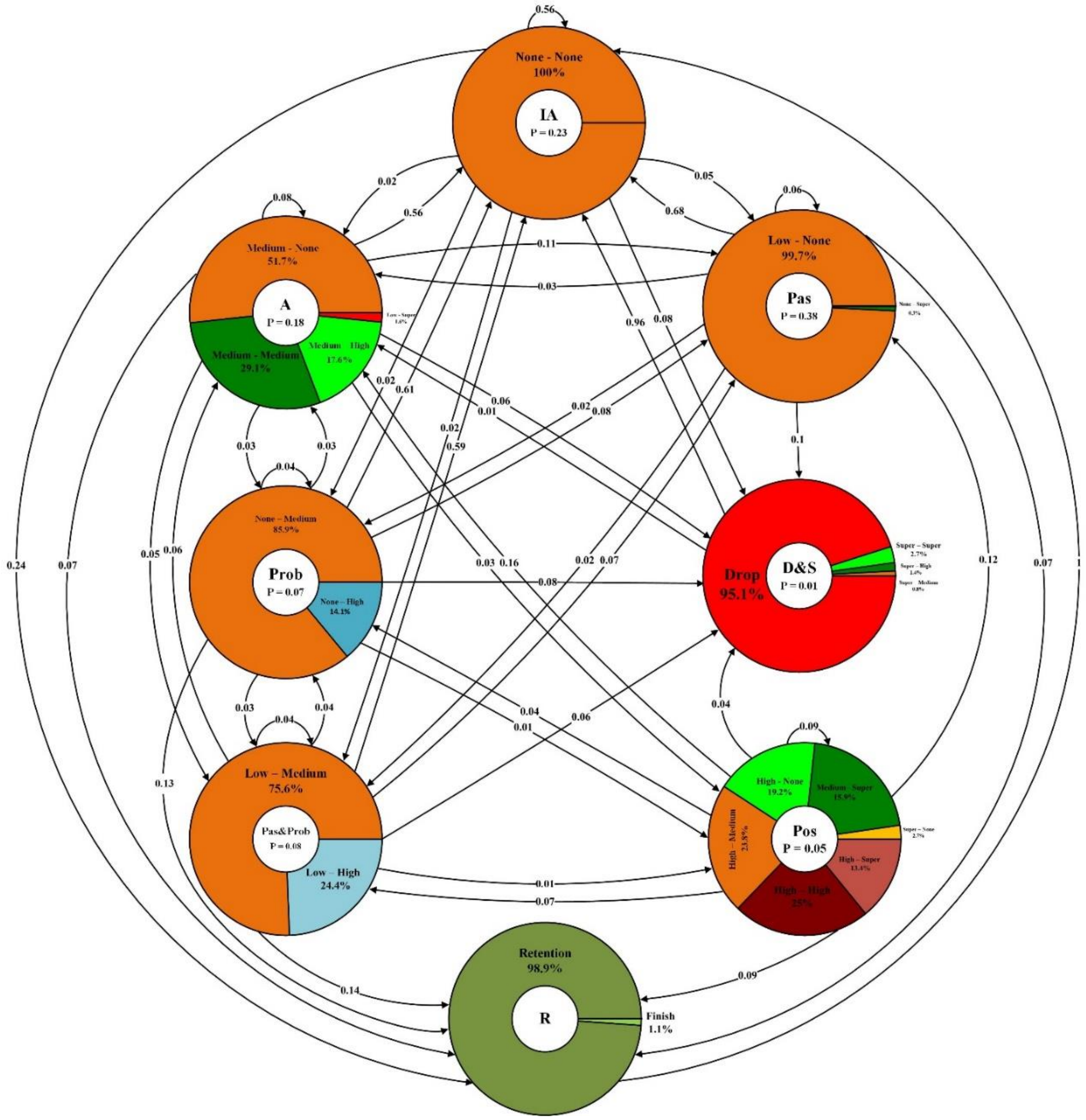


Figure 4.5. The latent state transition diagrams for early phase model

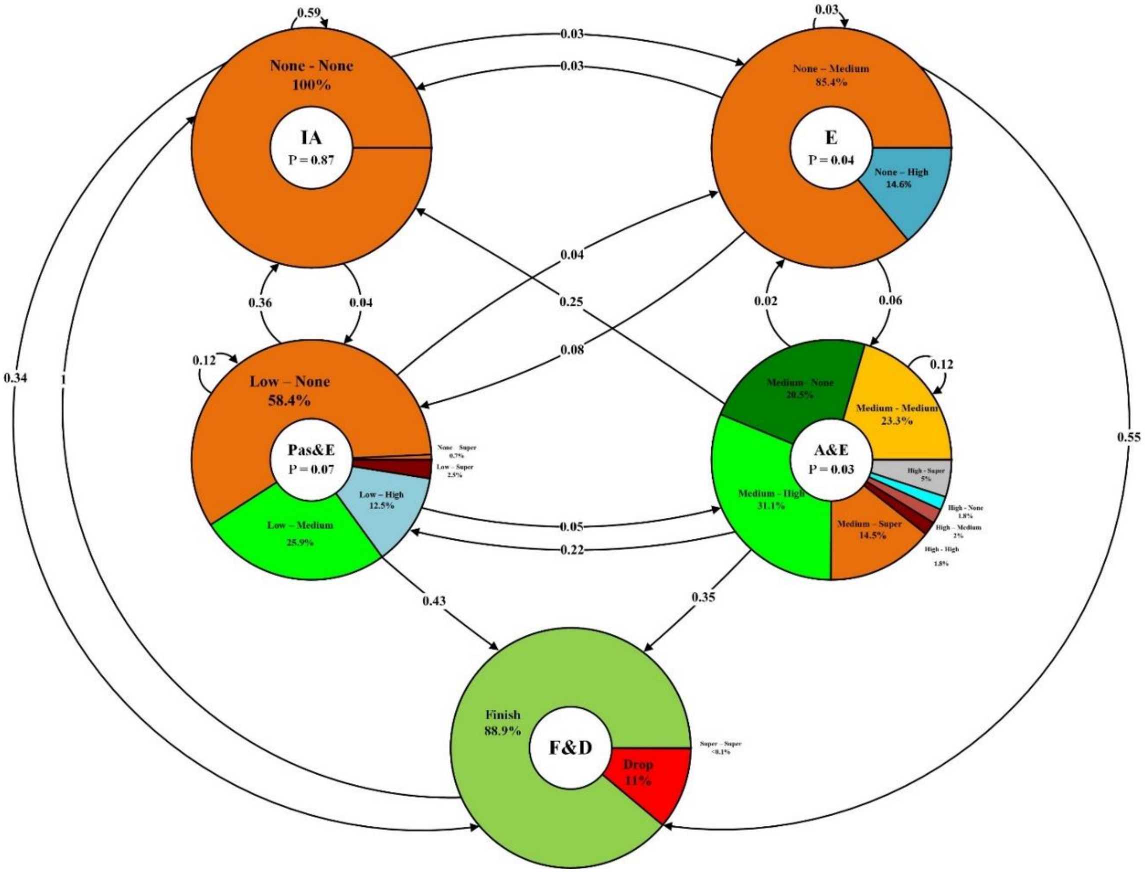
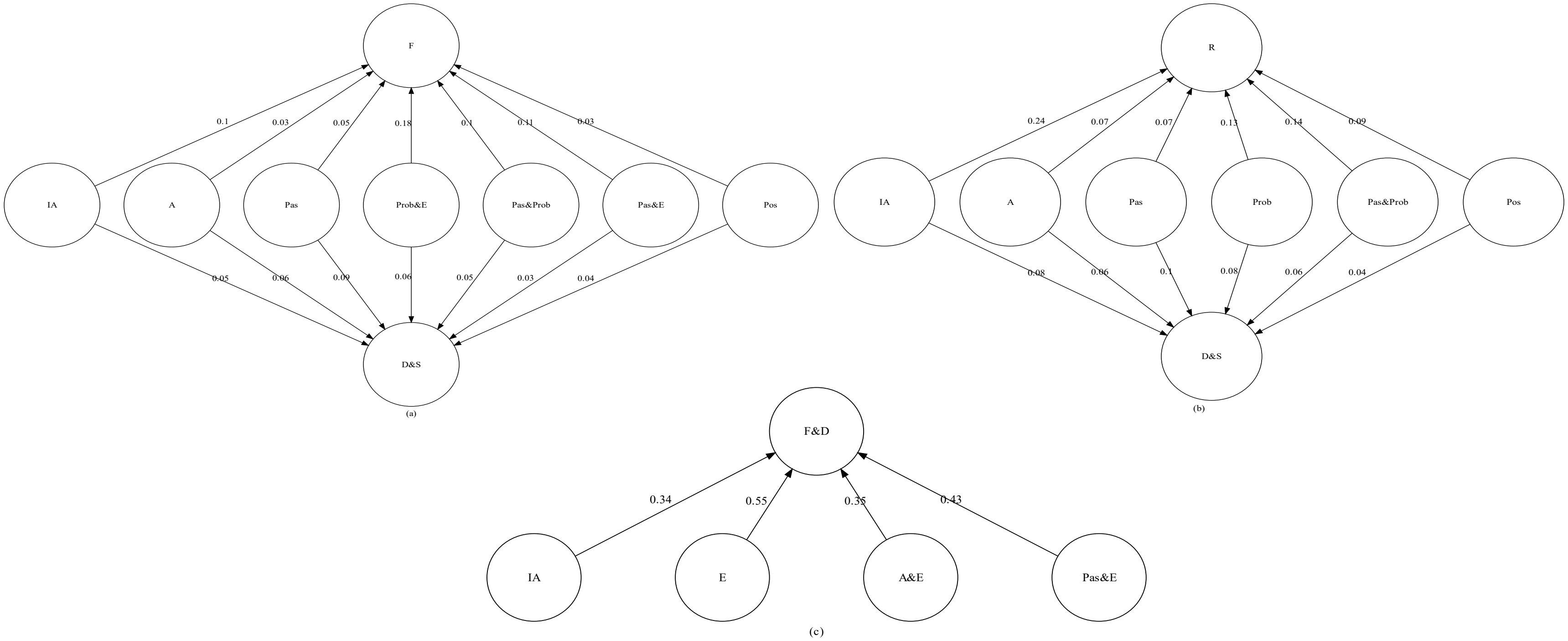


Figure 4.6. The latent state transition diagrams for late phase model



(a)

(b)

(c)

Figure. 4.7 Transitions to finish state and dropout state, (a) Seven-week model (b) Early phase model (c) Late phase model

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview of the Works

In this dissertation, we explore three important tasks in educational data mining. For the first problem, we explored the association between learners' forum participation and their performance. In this project, we first classified the MOOC learners' forum posts into topic-related and topic-unrelated. Then, we extracted the MOOC learners' longitudinal trajectory of online meaningful participation. The results of the project indicated the importance of learners' meaningful forum participation in their course grades and dropouts. The findings of this research provided significant implications on facilitating effective forum discussions and supporting learner performance in MOOCs. We further summarize our work into a journal paper titled "Untangling chaos in discussion forums: A temporal analysis of topic-relevant forum posts in MOOCs" (Yang et al., 2022)

The second project in this dissertation employed natural language processing and data mining techniques to explore the association between argumentative structure and essay scores. Our findings highlight the role of organizational argumentative structure in essay scoring. Furthermore, we found a common argumentative structure used by the high-scored essays. Finally, with the knowledge of argumentative elements and structures used in the previous essays, we proposed a feedback tool design to complement the current AES systems and help students improve their argument writing skill. Our findings can help students and educators to improve students' essay writing skills and support the automated essay assessment with detailed feedback.

Third, we explored the association between MOOC learners' forum behavior and dropout through multiple dimensions, including a quantitative dimension, a content dimension and a temporal dimension. To further investigate the prediction power of the MOOC learners' forum participation status, we built several prediction models based on some commonly used clickstream-based features and machine learning algorithms. Our findings show that applying longitudinal forum participation states as training features can improve the accuracy of the dropout prediction models. These results expand our understandings of learners' forum behavior and dropout and urge us to rethink the learners' forum behavior as a feature to predict their drop from multiple dimensions.

5.2 Future work

In this dissertation, we focused on how to apply educational data mining techniques to support and improve learners' learning experiences. For MOOC dropout analysis and prediction, further work could be done on including more specific forum topic categories, more MOOCs in different domains, and trying some deep learning models with larger datasets to achieve better performance. In addition, the patterns of MOOC learners' clickstream data can be further explored to understand the main causes of dropout and help reduce drop rate of MOOCs.

For argumentative essay writing feedback providing process, it is still a big challenge to consider all the features that could influence the essay score. Our work in this dissertation is only a small step towards a comprehensive AES feedback tool. For future work, the tool should take more features into consideration and provide comprehensive feedback to the students. We believe that our study can be further expanded to include other important features in essay writing to provide richer feedback to students and improve automatic essay scoring engine's performance. The possible features include the number of perspectives on the given prompt, the transitions between and within paragraphs, or the relevance of the examples to the prompt. In addition, it has been proved that automatic grading tools might have a grading bias for diverse students (Litman et al., 2021). Although our dataset contains essays from diverse populations, we did not analyze the structure difference between essays from culturally and linguistically diverse students. Further approaches could focus on analyzing the structure difference between diverse students' essays.

REFERENCES

Afrin, T., Wang, E., Litman, D., Matsumura, L. C., & Correnti, R. (2021). Annotation and classification of evidence and reasoning revisions in argumentative writing. *arXiv preprint arXiv:2107.06990*.

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips.

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Alrajhi, L., Alharbi, K., & Cristea, A. I. (2020, June). A multidimensional deep learner model of urgent instructor intervention need in MOOC forum posts. In *International conference on intelligent tutoring systems* (pp. 226-236). Springer, Cham.

Alhindi, T., & Ghosh, D. (2021). " Sharks are not the threat humans are": Argument Component Segmentation in School Student Essays. *arXiv preprint arXiv:2103.04518*.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), 537-553.

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

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z., & Wong, F. M. F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model. *IEEE transactions on Learning Technologies*, *7*(4), 346-359.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Boroujeni, M. S., Hecking, T., Hoppe, H. U., & Dillenbourg, P. (2017, March). Dynamics of MOOC discussion forums. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 128-137).

Boroujeni, M. S., & Dillenbourg, P. (2018, March). Discovery and temporal analysis of latent study patterns in MOOC interaction sequences. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 206-215).

Cabrio, E., & Villata, S. (2018, July). Five Years of Argument Mining: a Data-driven Analysis. In *IJCAI* (Vol. 18, pp. 5427-5433).

Capuano, N., & Caballé, S. (2019, November). Multi-attribute categorization of MOOC forum posts and applications to conversational agents. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing* (pp. 505-514). Springer, Cham.

Capuano, N., Caballé, S., Conesa, J., & Greco, A. (2021). Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 9977-9989.

Chakrabarty, T., Hidey, C., & McKeown, K. (2019). IMHO fine-tuning improves claim detection. *arXiv preprint arXiv:1905.07000*.

Chen, J., Feng, J., Sun, X., Wu, N., Yang, Z., & Chen, S. (2019). MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering, 2019*.

Chen, Y., & Zhang, M. (2017, May). MOOC student dropout: pattern and prevention. In *Proceedings of the ACM Turing 50th Celebration Conference-China* (pp. 1-6).

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219-243.

Crossley, S., Tian, Y., & Wan, Q. (2022). Argumentation Features and Essay Quality: Exploring Relationships and Incidence Counts. *Journal of Writing Research*.

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017, June). ReaderBench learns Dutch: building a comprehensive automated essay scoring system for Dutch language. In *International Conference on Artificial Intelligence in Education* (pp. 52-63). Springer, Cham.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, F., Zhang, Y., & Yang, J. (2017, August). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)* (pp. 153-162).

Diver, P., & Martinez, I. (2015). MOOCs as a massive research laboratory: Opportunities and challenges. *Distance Education*, *36*(1), 5-25.

Dumais, S. T. (2004). Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.*, *38*(1), 188-230.

Eriksson, T., Adawi, T., & Stöhr, C. (2017). "Time is the bottleneck": a qualitative study exploring why learners drop out of MOOCs. *Journal of Computing in Higher Education*, *29*(1), 133-146.

Engle, D., Mankoff, C., & Carbrey, J. (2015). Coursera's introductory human physiology course: Factors that characterize successful completion of a MOOC. *International Review of Research in Open and Distributed Learning*, *16*(2), 46-68.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015, March). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 146-150).

Fei, M., & Yeung, D. Y. (2015, November). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 256-263). IEEE.

Firmin, R., Schiorring, E., Whitmer, J., Willett, T., Collins, E. D., & Sujitparapitaya, S. (2014). Case study: Using MOOCs for conventional college coursework. *Distance Education*, *35*(2), 178-201.

Francois, J. M. (2006). Jahmm-An implementation of HMM in Java. *URL http://code. google. com/p/jahmm*.

Fournier-Viger, P., Gomariz, A., Gueniche, T., Mwamikazi, E., & Thomas, R. (2013, December). TKS: efficient mining of top-k sequential patterns. In *International Conference on Advanced Data Mining and Applications* (pp. 109-120). Springer, Berlin, Heidelberg.

Fournier-Viger, P., Lin, J. C. W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., & Lam, H. T. (2016, September). The SPMF open-source data mining library version 2. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 36-40). Springer, Cham.

Galikyan, I., Admiraal, W., & Kester, L. (2021). MOOC discussion forums: The interplay of the cognitive and the social. *Computers & Education*, *165*, 104133.

Gamage, D., Perera, I., & Fernando, S. (2020). Exploring MOOC user behaviors beyond platforms. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(8), 161-179.

Geigle, C., & Zhai, C. (2017, April). Modeling MOOC student behavior with two-layer hidden Markov models. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 205-208).

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*, 1-34.

Genolini, C., & Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, *25*(2), 317-328.

Ghosh, D., Khanam, A., Han, Y., & Muresan, S. (2016, August). Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 549-554).

Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *The Internet and Higher Education*, *23,* 18-26.

Goel, Y., & Goyal, R. (2020). On the effectiveness of self-training in mooc dropout prediction. *Open Computer Science, 10*(1), 246-258.

Goopio, J., & Cheung, C. (2021). The MOOC dropout phenomenon and retention strategies. *Journal of Teaching in Travel & Tourism*, *21*(2), 177-197.

Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of retention and achievement in a massive open online course. *American Educational Research Journal*, *52*(5), 925-955.

Handoko, E., Gronseth, S. L., McNeil, S. G., Bonk, C. J., & Robin, B. R. (2019). Goal setting and MOOC completion: A study on the role of self-regulated learning in student performance in massive open online courses. *International Review of Research in Open and Distributed Learning*, *20*(3).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*.

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014, March). Superposter behavior in MOOC forums. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 117-126).

Hussein, M. A., Hassan, H. A., & Nassef, M. (2020). A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, *11*(5).

Jeong, H., Biswas, G., Johnson, J., & Howard, L. (2010, June). Analysis of productive learning behaviors in a structured inquiry cycle using hidden Markov models. In *Educational Data Mining 2010*.

Jiang, Z., Zhang, Y., Liu, C., & Li, X. (2015). Influence Analysis by Heterogeneous Network in MOOC Forums: What Can We Discover?. *International Educational Data Mining Society*.

Jin, C. (2020). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments*, 1-19.

Ke, Z., & Ng, V. (2019, August). Automated Essay Scoring: A Survey of the State of the Art. In *IJCAI* (Vol. 19, pp. 6300-6308).

Kellogg, S., Booth, S., & Oliver, K. (2014). A social network perspective on peer supported learning in MOOCs for educators. *International Review of Research in Open and Distributed Learning*, *15*(5), 263-289.

Kellogg, S., & Edelmann, A. (2015). Massively open online course for educators (MOOC-E d) network dataset. *British Journal of Educational Technology*, *46*(5), 977-983.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing?. *arXiv preprint arXiv:1604.02416*.

Klebanov, B. B., Flor, M., & Gyawali, B. (2016, June). Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 63-72).

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs* (pp. 60-65).

Lalwani, A., & Agrawal, S. (2017). Few hundred parameters outperform few hundred thousand. In *Proceedings of the 10th International Conference on Educational Data Mining, EDM* (Vol. 17, pp. 448-453).

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, *45*(4), 765-818.

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of second language writing*, *33*, 21-34.

Lin, C., & Chi, M. (2017, June). A comparisons of bkt, rnn and lstm for learning gain prediction. In *International Conference on Artificial Intelligence in Education* (pp. 536-539). Springer, Cham.

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021, June). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. In *International Conference on Artificial Intelligence in Education* (pp. 255-267). Springer, Cham.

Mao, Y. (2018). Deep Learning vs. Bayesian Knowledge Tracing: Student Models for Interventions. *Journal of educational data mining*, *10*(2).

Mayfield, E., & Black, A. W. (2020, July). Should you fine-tune BERT for automated essay scoring?. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 151-162).

Mim, F. S., Inoue, N., Reisert, P., Ouchi, H., & Inui, K. (2019, July). Unsupervised learning of discourse-aware text representation for essay scoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 378-385).

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003, November). Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003.* (Vol. 1, pp. T2A-13). IEEE.

Molenaar, I. (2014). Advances in Temporal Analysis in Learning and Instruction. *Frontline Learning Research*, *2*(4), 15-24.

Moreno-Marcos, P. M., Muñoz-Merino, P. J., Maldonado-Mahauad, J., Perez-Sanagustin, M., Alario-Hoyos, C., & Kloos, C. D. (2020). Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. *Computers & Education*, *145*, 103728.

Nagata, R. (2019, November). Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3206-3215).

Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017, April). MOOC dropout prediction: lessons learned from making pipelines interpretable. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 351-359).

Nguyen, H., & Litman, D. (2018, April). Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Niculae, V., Park, J., & Cardie, C. (2017). Argument mining with structured SVMs and RNNs. *arXiv preprint arXiv:1704.06869*.

Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017, March). A neural network approach for students' performance prediction. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 598-599).

OpenNLP, A. (2011). Apache software foundation. *URL http://opennlp. apache. org*.

Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers & Education*, *143*, 103673.

Persing, I., & Ng, V. (2015, July). Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 543-552).

Persing, I., & Ng, V. (2016, June). End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1384-1394).

Pessoa, S., Mitchell, T. D., & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, *38*, 42-55.

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015, September). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 431-439).

Pillutla, V. S., Tawfik, A. A., & Giabbanelli, P. J. (2020). Detecting the depth and progression of learning in massive open online courses by mining discussion data. *Technology, Knowledge and Learning*, *25*(4), 881-898.

Pursel, B. K., Zhang, L., Jablokow, K. W., Choi, G. W., & Velegol, D. (2016). Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning*, *32*(3), 202-217.

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education, 27*(4), 694-728.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Ray, S., & Turi, R. H. (1999, December). Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques* (Vol. 137, p. 143).

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Semenova, T. (2022). The role of learners' motivation in MOOC completion. *Open Learning: The Journal of Open, Distance and e-Learning*, *37*(3), 273-287.

Speck, J., Gualtieri, E., Naik, G., Nguyen, T., Cheung, K., Alexander, L., & Fenske, D. (2014, March). ForumDash: Analyzing online discussion forums. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 139-140).

Stephens-Martinez, K., Hearst, M. A., & Fox, A. (2014, March). Monitoring moocs: which information sources do instructors value?. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 79-88).

Stab, C., & Gurevych, I. (2014, October). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 46-56).

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2019). Understanding teacher shortages: An analysis of teacher supply and demand in the United States. *Education Policy Analysis Archives, 27*(35).

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia computer science*, *169*, 726-743.

Taghipour, K., & Ng, H. T. (2016, November). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891).

Talavera, L., & Gaudioso, E. (2004, August). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence* (pp. 17-23).

Tang, H., Xing, W., & Pei, B. (2018). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education*, *39*(3), 353-372.

Tang, C., Lau, R. W., Li, Q., Yin, H., Li, T., & Kilis, D. (2000, June). Personalized courseware construction based on web data mining. In *Proceedings of the first international conference on web information systems engineering* (Vol. 2, pp. 204-211). IEEE.

Tang, S., Peterson, J. C., & Pardos, Z. A. (2016, April). Deep neural networks and how they apply to sequential education data. In *Proceedings of the third (2016) acm conference on learning@ scale* (pp. 321-324).

Tang, X., Li, S., & Huang, Z. (2020, December). The relationship between mode and content type of forum interaction and MOOC engagement pattern. In *2020 Ninth International Conference of Educational Innovation through Technology (EITT)* (pp. 182-187). IEEE.

Tsantis, L., & Castellani, J. (2001). Enhancing learning environments through solution-based knowledge discovery tools: Forecasting for self-perpetuating systemic reform. *Journal of Special Education Technology*, *16*(4), 39-52.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*(2), 459-484.

Uto, M., Xie, Y., & Ueno, M. (2020, December). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6077-6088).

Wachsmuth, H., Al Khatib, K., & Stein, B. (2016, December). Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers* (pp. 1680-1691).

Wan, Q., Crossley, S., Banawan, M., Balyan, R., Tian, Y., McNamara, D., & Allen, L. (2021). Automated Claim Identification Using NLP Features in Student Argumentative Essays. *International Educational Data Mining Society*.

Wang, H., Huang, Z., Dou, Y., & Hong, Y. (2020, December). Argumentation mining on essays at multi scales. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5480-5493).

Wang, W., Yu, H., & Miao, C. (2017, July). Deep model for dropout prediction in MOOCs. In *Proceedings of the 2nd international conference on crowd science and engineering* (pp. 26-32).

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society*.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). Delving deeper into MOOC student dropout prediction. *arXiv preprint arXiv:1702.06404*.

Whitehill, J., Williams, J., Lopez, G., Coleman, C., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout. *Available at SSRN 2611750*.

Wilson, K. H., Xiong, X., Khajah, M., Lindsey, R. V., Zhao, S., Karklin, Y., ... & Mozer, M. C. (2016). Estimating student proficiency: Deep learning is not the panacea. In *In Neural Information Processing Systems, Workshop on Machine Learning for Education* (Vol. 3).

Wise, A. F., Cui, Y., Jin, W., & Vytasek, J. (2017). Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, 32, 11-28.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). The WEKA workbench. online appendix for “Data Mining: Practical machine learning tools and techniques”. In *Morgan Kaufmann*.

Xue, J., Tang, X., & Zheng, L. (2021). A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring. *IEEE Access*, *9*, 125403-125415.

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, *57*(3), 547-570.

Xing, W., Tang, H., & Pei, B. (2019). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *The Internet and Higher Education*, *43*, 100690.

Xiong, X., Zhao, S., Van Inwegen, E. G., & Beck, J. E. (2016). Going deeper with deep knowledge tracing. *International Educational Data Mining Society*.

Yang, B., Tang, H., Hao, L., & Rose, J. R. (2022). Untangling chaos in discussion forums: A temporal analysis of topic-relevant forum posts in MOOCs. *Computers & Education*, *178*, 104402.

Yang, D., Piergallini, M., Howley, I., & Rose, C. (2014). Forum thread recommendation for massive open online courses. In *Educational Data Mining 2014*.

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020, November). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560-1569).

Yang, Y., & Zhong, J. (2021, May). Automated essay scoring via example-based learning. In *International Conference on Web Engineering* (pp. 201-208). Springer, Cham.

Yannakoudakis, H., & Briscoe, T. (2012, June). Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 33-43).

Ye, X., & Manoharan, S. (2019, December). Providing automated grading and personalized feedback. In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing* (pp. 1-5).

Yeung, C. K., & Yeung, D. Y. (2018, June). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1-10).

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015, June). Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 224-232).

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L. C., ... & Quintana, R. (2019, July). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9619-9625).

Zheng, Y., Gao, Z., Wang, Y., & Fu, Q. (2020). MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series. *IEEE Access*, *8*, 225324-225335