Unsupervised Modeling for Understanding MOOC Discussion Forums: A Learning Analytics Approach

Aysu Ezen-Can
Department of Computer Science
North Carolina State University
aezen@ncsu.edu

Kristy Elizabeth Boyer
Department of Computer Science
North Carolina State University
keboyer@ncsu.edu

Shaun Kellogg
Friday Institute for Educational Innovation
North Carolina State University
sbkellog@ncsu.edu

Sherry Booth
Friday Institute for Educational Innovation
North Carolina State University
sherry_booth@ncsu.edu

ABSTRACT

Massively Open Online Courses (MOOCs) have gained attention recently because of their great potential to reach learners. Substantial empirical study has focused on student persistence and their interactions with the course materials. However, most MOOCs include a rich textual dialogue forum, and these textual interactions are largely unexplored. Automatically understanding the nature of discussion forum posts holds great promise for providing adaptive support to individual students and to collaborative groups. This paper presents a study that applies unsupervised student understanding models originally developed for synchronous tutorial dialogue to MOOC forums. We use a clustering approach to group similar posts, compare the clusters with manual annotations by MOOC researchers, and further investigate clusters qualitatively. This paper constitutes a step toward applying unsupervised models to asynchronous communication, which can enable massive-scale automated discourse analysis and mining to better support students' learning.

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education; K.3.1 [**Computers and Education**]: Distance learning—MOOC

General Terms

Human Factors

Keywords

Text-based learning analytics, MOOCs, Online Learning

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA.

http://dx.doi.org/10.1145/2723576.2723589 ...$15.00.

1. INTRODUCTION

Recent years have witnessed a growing body of learning analytics research on MOOCs including visualizing and analyzing patterns of student engagement [1, 5], understanding factors related to disengagement [13], scaling up learning analytics [11], students' performance analysis [19], exploration of participation [4] and automated student modeling [24]. These studies have highlighted important dynamics of MOOCs in terms of learner activities and participation.

One very important source of data in MOOCs is the textual dialogue among students and course support staff on MOOC discussion forums. Discussion forums can be utilized for a wide variety of purposes, chief among these for learners to help each other and to discuss thier viewpoint on course topics [23]. Compared to performance analysis of students in MOOCs, research on these rich sources of textual data is more limited. Sentiment analysis in MOOC discussion forums [22], identifying exploratory dialogue [8] and utilizing topic models for predicting student survival [16] have been shown to be promising. Highlighting the importance of forum posts, the effect of number of posts on a student's final grade has been shown [15]. However, there is still much to be explored in terms of textual information in online courses, particularly regarding how the pragmatics of dialogue unfold in these asynchronous discussion forums and how those dialogues support learning.

In order to model discussion forum dialogues, one important level of modeling is *dialogue acts*, which can be thought of as summarizing the pragmatic role of each utterance, such as asking a question, agreeing or disagreeing, and providing feedback. Our recent work has developed an unsupervised machine-learning approach to this problem [6], improving upon earlier reported unsupervised classifiers for dialogue [18]. The great promise of unsupervised models for this task is reinforced by the strong performance of unsupervised models for other learning analytics goals such as modeling learner subpopulations [13] and refining automatic ITS capabilities [21].

This paper brings together an unsupervised dialogue act classification framework with MOOC modeling approaches, with the primary goal of gaining insights about the structure of forum posts in a MOOC. We empirically investigate the performance of an unsupervised dialogue act classification model built for synchronous tutorial dialogue on the

asynchronous MOOC discussion data. These two types of dialogues are different in structure in several ways, yet the results demonstrate the model's cross-domain applicability.

2. CORPUS

The MOOC discussion forum data investigated in this study were collected within an 8-week MOOC for Educators (MOOC-Ed) titled *Planning for the Digital Learning Transition in K-12 Schools.* Enrolled learners within the MOOC were teachers already practicing within K-12 environments. The MOOC was held in fall of 2013. The course was designed to help school and district leaders plan and implement K-12 digital learning initiatives. In support of this goal, ongoing opportunities for professional collaboration were provided through the discussion forum, which was utilized for online discussions, peer feedback on project submissions, and for sharing resources and information. The corpus of discussions includes 550 posts from 155 learners, totaling 57 distinct discussions. The number of posts written per learner throughout the course varied from a minimum of 1 to a maximum of 22.

In earlier work, the content of communications exchanged among peers was manually annotated in order to better understand the purpose of each forum post and products of the exchanges [12]. The manual annotations adopted a simplified version of Henri's model [10] for content analysis of computer-mediated communication to assess the extent of critical thinking demonstrated by individual participants (not applicable, undeveloped thinking, illogical thinking, critical thinking), as well the Interaction Analysis Model [9] to determine the extent to which new knowledge was collectively constructed (sharing/comparing information, discovery and exploration of ideas, negotiation of meaning, testing and modification of synthesis). Aside from assessing instructional outcomes, one goal of combining content and network analyses was to explore the relationship between network structures and these interaction outcomes to inform design of future MOOC-Eds. In the current work, we use these criteria to compare the groupings learned by the unsupervised model.

3. EXPERIMENTS

The goal of the experiments is to empirically investigate the extent to which unsupervised models can provide insights into the flow of conversations among learners on a discussion forum. The unsupervised modeling approach is a fully data-driven approach, with its machine-learned models not trained on any manually labeled data. With this model in hand we then compare the automatically learned clusters against the distribution of manual labels (from the prior analysis) in those clusters. We further investigate the clusters qualitatively to gain deeper insight into their structure. Finally we combine the clustering process with automatic topic modeling to provide a rich understanding of discussion posts in the MOOC.

3.1 Clustering Algorithm

The unsupervised machine learner that is utilized for this task is the k-medoids clustering algorithm with a greedy seed selection approach using bag-of-words, which has been shown to be successful for dialogue act classification with synchronous tutorial dialogue in our prior work [7]. We utilize each forum post as an utterance for the k-medoids algorithm. k-medoids clustering is a well-known clustering technique that takes actual data points as the center of each cluster [14], in contrast to the perhaps better-known technique of k-means in which the center of each cluster is obtained via averaging and therefore is likely not a data point in the corpus. k-medoids requires initial seeds for clusters in its first iteration, and to provide better seeding in the first iteration, we use a greedy seed selection approach similar to the one used in k-means++ [2] which selects the first seed randomly and then greedily chooses seeds that are farthest from the chosen seeds.

Determining number of clusters. In order to proceed with unsupervised modeling of the discussion forum corpus into clusters, we first determine the most suitable number of clusters. Determining this parameter is known to be a challenging issue in clustering and can be fraught with subjectivity. For the current investigation we rely on the Bayesian Information Criterion (BIC), a metric that considers model fit and penalizes for increased model complexity (higher number of clusters). Smaller BIC values are preferred, and the minimum BIC value for this dataset was achieved at k=7 clusters (Figure 1). Seven clusters was also the knee in the inter-cluster distance graph as shown in Figure 2. Inter-cluster distance represents the distances between clusters, with higher values being preferred.

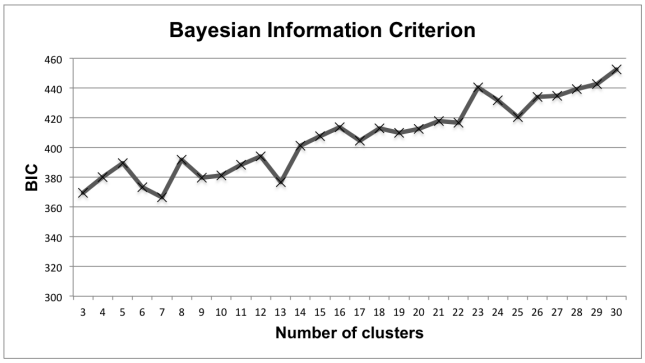

Figure 1: BIC values for varying number of clusters.

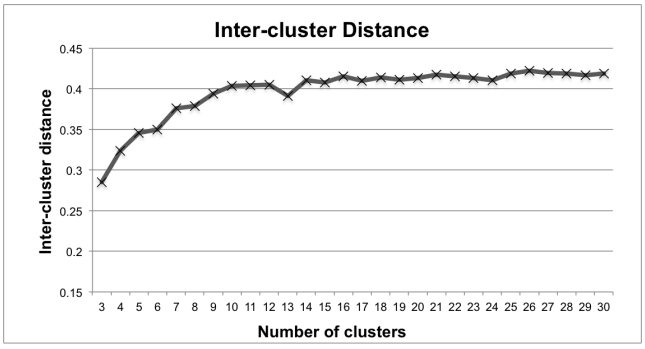

Figure 2: Average inter-cluster distances for varying number of clusters.

3.2 Clustering Results

With the best-fit unsupervised clustering in hand, we move on to analyzing the model, which has grouped each discussion forum post into one of seven clusters. We start our investigation by exploring the distribution of discussion post types,

as manually labeled in prior work, over the clusters. As depicted in Figure 3, statements are very frequent, appearing in all of the clusters. To distinguish statements further, we also conducted qualitative inspection which is explained in the Section 3.3. Cluster 5 is different from other clusters in that it contains more scaffolding posts, and in Cluster 6, we see many questioning posts.

Moving further in the analysis, the second criterion we examine is the extent to which the clusters align with the manually annotated "evolution of ideas". We see that the cluster that has most of the scaffolding posts (Cluster 5) is mostly composed of information sharing and comparison of ideas. In addition, this cluster is also the cluster that has the highest percentage of "undeveloped thinking" according to the manual labeling.

Clusters 1, 2 and 3, which have many statements, also have a large number of posts that are manually labeled as illogical thinking. Here it is important to note that according to the manual annotation scheme, there are multiple labels assigned to each post; this structure influences the extent to which a single label should be "assigned" to each automatically extracted cluster.

As the graphs have shown, the clusters are not pure compared to any dimension of the hand-crafted labels. However, the unsupervised model grouped these clusters for a reason: they were structurally similar in a systematic way that was evident from the data but different from the criteria used to develop the manual tagging scheme. To begin to discover what this structure is, we investigate the clusters qualitatively.

3.3 Qualitative Evaluation of Clusters

Table 1 shows sample posts from each cluster. Cluster 1 captures *agreement*, such as "Great point..." and "I appreciate..." Cluster 2 expresses personal viewpoints that emphasize the importance of topics raised in the course, such as "I'm excited about..." and "...plays a huge role..." In Cluster 3 declarative statements, mostly factual in nature, are grouped together. Cluster 4 is characterized by questions and disagreements, for example "PD is important but..." and "How are your schools and districts using..." Cluster 5 is composed of utterances that express appreciation of others' ideas, such as "I love your idea..." and "Well thought out plan." Cluster 6 features questions that the learners pose to each other. Finally, Cluster 7 groups posts in the form of statements which evaluate the current education system and compare it to the future, such as "We need to move toward..." and "Teachers are no longer the only..."

3.4 Topic Modeling

Inspecting the clusters in detail reveals that they capture not only dialogue acts such as questions, declarative statements, and agreements, but they also capture what can be considered an orthogonal dimension: the *topic* of the discussions. This finding has been observed in other unsupervised modeling of dialogue data where the same dialogue act and topic interplay was observed for Twitter [17] and in our own tutorial dialogue work [6]. Unlike in some other types of (particularly synchronous) dialogue where the topics are more systematic and constrained, in online discussions the topics range widely. These topics may be very important for understanding and providing automated support to learners [20]. To extract topics we applied Latent Dirichlet Allocation, an automatic topic modeling technique [3] to capture the topical theme of each cluster and assumed that each post had a different topic. Table 1 shows top topic words/ phrases of each cluster. The results show that Cluster 1 is not only typically agreements, but they tend to be about resources and student achievement. Cluster 2 not only expresses personal ideas but their topics tend to compare the future to the past. Cluster 3 contains factual moves, and more specifically these tend to be about accessing emerging types of resources. Cluster 4 groups disagreements and questions on classrooms, tools and programs. Cluster 5 expresses appreciation about proposed plans. The questions in Cluster 6 are mostly related to student learning, whereas Cluster 7 has statements that appear to be characterized by topics on support and equity.

4. CONCLUSION

Massively Open Online Courses bring education opportunities to huge audiences. Understanding how students learn and collaborate within MOOCs is a tremendous opportunity for the learning analytics community, and models built upon MOOC data hold the potential to provide adaptive real-time support to learners, improving their experience. This paper has presented an empirical investigation of the textual discussion forum data from a MOOC, with the goal of automatically extracting the structure of the discussions to understand students posts better.

The results of this investigation indicate that it is possible to apply an unsupervised modeling framework developed for synchronous conversations to asynchronous discussions, although important differences in structure suggest it is necessary to perform an additional step of topic modeling in order to form more interpretable and cohesive models. Unsupervised modeling techniques can provide insights into potentially huge numbers of similar posts and the topics that learners are talking about. In the future, real-time remedial support can be facilitated by these automated models, and then provided by course staff or by intelligent support systems, to provide an adaptive and highly effective learning experience on a large scale.

5. ACKNOWLEDGMENTS

The authors wish to thank the members of the Center for Educational Informatics at North Carolina State University for their helpful input on this work. The MOOC was hosted in partnership with the Alliance for Excellent Education and was funded by the Gates Foundation and led by Athabasca University as part of the MOOC Research Initiative.

6. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 687–698, 2014.

[2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] D. Clow. MOOCs and the funnel of participation. In *Proceedings of LAK*, pages 185–189, 2013.

Cluster #	Cluster Contents	LDA Topic Words
1	- I also agree PERSON that without more PD putting technology into student hands will not benefit the students (...) - Great point. When I read the discussion for this week I thought of the same thing (...) - I have to agree whole-heartedly with PERSON ... professional development is key. (...) - I appreciate so much your thinking on this issue. (...) - What a well-thought out plan . (...)	find millions, students' parents, agree person, earn minimum, valuable resources, entire district, technology initiatives, failing student
2	- I'm excited about getting technology into the hands of our learners (...) - Gaining insight into the transformation of education for future learners is truly a direction (...) - I think social media plays a huge role in what how when and where students learn. (...) - I think the most challenging and critical role that the teacher will play (...)	st century, byod program, high school, compel today, digital learning, greatest experiences, social media, foot curly cords, skill sets, technology integration, hard pressed, years ago
3	- Right now in my school I made a project (...) - My son just started kindergarten this year and over the summer (...) - Not everybody learns best using paper and pencil. (...) - I am often left wondering what is the role of direct instruction (...)	symbaloo pages access, linking resources, digital learning, primary sources, wide step, cell phones, student test, complete digital learning lessons, student test scores
4	- How are your schools and districts using different strategies (...) - If digital learning is implemented effectively what will be different for students and teachers? - PD is important but (...) - The Utopia is for people that want to live in it!! (...)	district supported, technology tools, student tech program, school technology, digital learning, professional development, integrates eportfolios assessments
5	- That's too good that you have taken initiative to implement 1:1 computing. (...) - I like that you included student assessments and project based learning. (...) - I also love your idea of an online help desk. (...) - Great plan for the tech day workshops. - Well thought out plan.	valuable resource, rural farming community, interesting idea, agree robotics, problem solving, pd initiative, great plan, big undertaking, imagine students, included student assessments
6	- How would you like to expand or have you expanded student learning opportunities? - How do you think we should change how when and where students learn? - How do you think community partnerships or internships could affect (...) - What role do you think social media plays	enhance students learning, internships provide valuable, students learn, community partnerships state curriculum, students learn, equal footing, marketable skills
7	- The role of the teacher will be so much different. (...) - Traditional teachers such as myself will have to learn whole new techniques (...) - We need to move toward a model of education (...) - Teachers are no longer the only holders of knowledge. (...)	staff students parents, school districts, find resources, elementary education folks, successful programs, learning process, inevitable support issues, high poverty

Table 1: Sample forum posts from each cluster and their top topic words. The 'PERSON' tag is used to remove names.

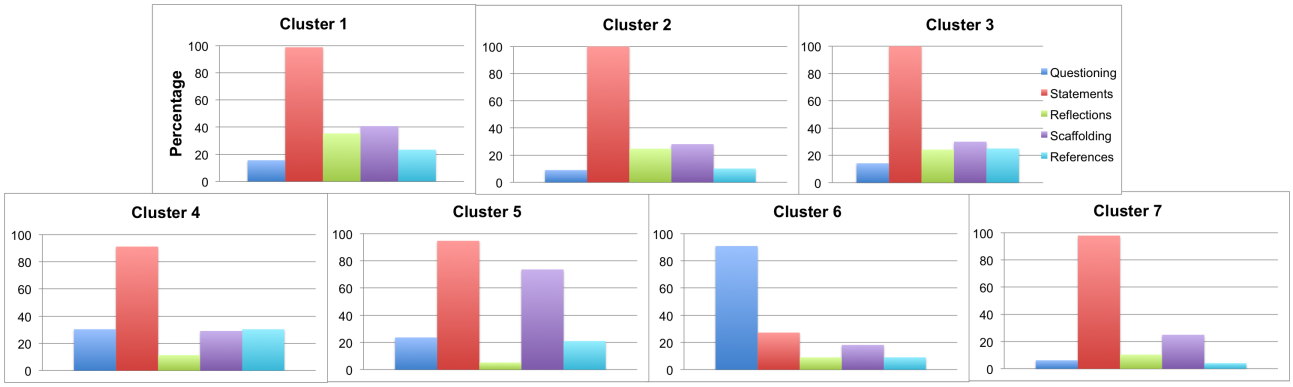

Figure 3: Distribution of purposes of forum posts among clusters.

[5] C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy. Visualizing patterns of student engagement and performance in MOOCs. pages 83–92, New York, New York, USA, 2014. ACM Press.

[6] A. Ezen-Can and K. E. Boyer. Unsupervised classification of student dialogue acts with query-likelihood clustering. In *Proceedings of EDM*, pages 20–27, 2013.

[7] A. Ezen-Can and K. E. Boyer. Toward adaptive unsupervised dialogue act classification in tutoring by gender and self-efficacy. In *Extended Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, pages 94–100, 2014.

[8] R. Ferguson, Z. Wei, Y. He, and S. Buckingham Shum. An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of LAK*, pages 85–93, 2013.

[9] C. N. Gunawardena, C. A. Lowe, and T. Anderson. Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, 17(4):397–431, 1997.

[10] F. Henri. Computer conferencing and content analysis. In *Collaborative learning through computer conferencing*, pages 117–136, 1992.

[11] D. T. Hickey, T. A. Kelley, and X. Shen. Small to big before massive: scaling up participatory learning analytics. In *Proceedings of LAK*, pages 93–97, 2014.

[12] S. B. Kellogg, S. Booth, and K. M. Oliver. A social network perspective on peer support learning in MOOCs for educators. In *International Review of Research in Open and Distance Learning*, in press.

[13] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of LAK*, pages 170–179, 2013.

[14] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 144–155, 1994.

[15] S. Palmer, D. Holt, and S. Bray. Does the discussion help? The impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology*, 39(5):847–858, 2008.

[16] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Understanding MOOC discussion forums using seeded LDA. In *Proceedings of the 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[17] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of Twitter conversations. In *Proceedings of the Association for Computational Linguistics*, pages 172–180, 2010.

[18] V. Rus, C. Moldovan, N. Niraula, and A. C. Graesser. Automated discovery of speech act categories in educational games. In *Proceedings of EDM*, pages 25–32, 2012.

[19] J. L. Santos, J. Klerkx, E. Duval, D. Gago, and L. Rodríguez. Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses. In *Proceedings of LAK*, pages 98–102, 2014.

[20] S. Shatnawi, M. M. Gaber, and M. Cocea. Automatic content related feedback for MOOCs based on course domain ontology. In *Intelligent Data Engineering and Automated Learning–IDEAL 2014*, pages 27–35. 2014.

[21] J. Stamper and T. Barnes. Unsupervised MDP Value Selection for Automating ITS Capabilities. In *Proceedings of EDM*, pages 180–189, 2009.

[22] M. Wen, D. Yang, and C. P. Rosé. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In *Proceedings of EDM*, pages 130–137, 2014.

[23] J. Yoo and J. Kim. Capturing Difficulty Expressions in Student Online Q&A Discussions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 208–214, 2014.

[24] M. Yudelson, R. Hosseini, A. Vihavainen, and P. Brusilovsky. Investigating automated student modeling in a Java MOOC. In *Proceedings of EDM*, 2014.