Topic Evolution Models for Long-running MOOCs

Arti Ramesh[1] and Lise Getoor[2]
SUNY Binghamton[1] University of California, Santa Cruz[2]
artir@binghamton.edu, getoor@ucsc.edu

Abstract. Massive open online courses (MOOCs) have emerged as a powerful platform for imparting education in the last few years. Discussion forums in online courses connect various geographically separated MOOC participants and serve as the primary means of communication between them. The text in the forums reflects many important aspects of the course such as the student population and their changing interests, parts of the course that were well received and parts needing attention, and common misconceptions faced by students. In order to improve the quality of online courses and students' interaction and learning experience, instructors need to actively monitor and discern patterns in previous iterations of the course and mold the course to suit the needs of the ever-changing student population. To enable this, in this work, we perform a systematic detailed analysis of the evolution of fine-grained topics in online course discussion forums across repeated MOOC offerings using seeded topic models and draw important insights on the nature of students, types of issues, and student satisfaction. We present topic evolution results on two successful long-running MOOCs: i) a business course, and ii) a computer science course. Our models uncover interesting topic trends in both courses including the decline of logistic issues in both courses as iterations unfold, decline in grading related issues when automatic grading is adopted in the business course, and prevalence of technical issues in the computer science course in comparison to the business course. Our models throw light on the different ways students interact on MOOCs and their changing needs, and are useful for instructors to understand the progression of courses and accordingly fine-tune courses to meet student expectations.

1 Introduction

Massive open online courses (MOOCs) are increasingly becoming a powerful educational platform, providing students from all over the world access to high quality education. As MOOCs continue to grow, instructors are faced with the problem of understanding the needs and expectations of the ever-changing student population, molding the course to better suit their interests, identifying issues in past iterations, and addressing them in future iterations. This endeavor ensures a smoother delivery of the course and helps in fostering a superior learning experience.

With MOOCs, there is tremendous opportunity to develop models that automatically gauge feedback by interpreting textual content in the discussion forums. The text in the forums reflects many important aspects of the course such as the student population and their changing interests, parts of the course that were well received and parts needing attention, and common misconceptions faced by students. In the existing framework, online instructors often resort to manually poring through posts in the

forums to determine improvements for future iterations. And, when they make improvements, they again have to rely on manual effort to determine whether the improvements have been helpful. An automatic way to quantitatively measure how the health of the course and the interests of the student population are evolving over time will be helpful for instructors to mold their course to fit student needs better and help improve their online interaction and learning experience. While most previous work in this space interpret text in the discussion forums of individual courses [5,6,13,11,4], analyzing text across repeated offerings of a course provides a panoramic view of course progression. This analysis can potentially help instructors discern topic patterns corresponding to relevant topics such as course materials and issues, and focus limited instructor resources on addressing the most prevalent and relevant problems.

In this work, we develop weakly-supervised topic evolution models to analyze posts in online course discussion forums across repeated offerings. We leverage a seeded variant of topic modeling, seeded latent dirichlet allocation (seeded LDA) [9], to induce and track evolution of specific topic clusters relevant to online courses across iterations. The large number of posts in MOOC discussion forums in each iteration of the course and privacy issues surrounding the creation and distribution of labeled data make weakly-supervised seeded topic evolution models attractive for understanding forum posts and improving future course offerings.

We present a fine-grained analysis of discussion forums in two successful and popular MOOCs that have run for several iterations. We first identify important topic themes relevant in online courses. Next, we track the rise and decline of these topics as the iterations unfold. This analysis reveals how the focus of the course, the student population, and their needs are evolving with time. Our models and analysis are useful for educators, MOOC practitioners, and instructors to understand the longitudinal evolution of online courses. Our analysis is also helpful to instructors to fine-tune their courses to meet student expectations, which subsequently helps in achieving superior interaction and learning experience, when students interact on the MOOC. To the best of our knowledge, ours is the first work that models evolution of topics across repeated offerings of online courses.

Our main contributions in this work are as follows:

- We show how to use seeded LDA to categorize discussion forum posts in online courses. We construct *four* different seeded LDA models for each of the two courses using a common set of seed words. Using our models, we track the progression of seeded topics and draw important insights on topic patterns of two successful long-running MOOCs: i) *thirty four* iterations of a business course (BUSINESS), and 2) *fifteen* iterations of a computer science course (CS).
- We identify the three primary purposes of forums in online courses upon carefully mining posts across different courses: i) socializing with fellow classmates (*social*), ii) reporting issues in the course (*issue*), and iii) discussing course related material (*technical*). We categorize the posts according to these three primary purposes of the forums. Our temporal analysis uncovers changes in topic patterns such as an increase in issue posts after the 4^{th} iteration in the CS course, when the course splits into two parts. Our temporal analysis is helpful in understanding how big course changes such as splitting the course affects the students.

- Next, we categorize the posts referring to the three most important *course elements* in online courses: i) lectures, ii) quizzes, and iii) certificate, to understand the emphasis on each of them, respectively. While lectures are the most popular course element in the BUSINESS course, quizzes surpass lectures in the CS course. We also observe that certificate receives more attention in the BUSINESS course in comparison to the CS course. This analysis provides insight on which course elements get more attention in the course and subsequently, the nature of students in both the courses, and their evolution with time.
- We then show how to use a combination of seeded topic models to perform a finer-grained analysis of issue posts and study their distribution across: i) lecture, quiz, and certificate course elements, and ii) fine-grained lecture and quiz sub-topics. We find that though lectures are the most dominant course element in the BUSINESS course, most issues are reported on quizzes. We also observe that while grading issues tend to occur across both the courses, submission issues predominate the forums in the CS course. Interestingly, we notice that grading issues decline in the BUSINESS course after peer grading is replaced by automatic grading, indicating a general preference for the latter. Our fine-grained analysis sheds light on issues faced by students across iterations that could negatively impact student satisfaction and enrollment in future iterations.
- In the CS course, we observe another important dimension, *technical* posts, dominating the issue posts. Technical issues in software installation and code compilation are unique to computer science courses. We use a combination of seeded LDA models to separate logistic issues from technical issues and find that logistic issues follow a similar trend to the BUSINESS course, declining with time, indicating the need to focus on technical issues in the CS course.

2 Related Work

Recently, there has been a growing interest in understanding text in online course discussion forums. Previous work in this space focus on understanding forum content in individual courses to improve the experience of MOOC participants [5,6,13,11,1,4]. To the best of our knowledge, ours is the first work studying the evolution of forum content in online courses over time. Topic evolution has been studied previously on scientific research papers [8,7]. Temporal variants of LDA such as Dynamic Topic Model (DTM) [3] and Topics over Time (ToT) [12] model LDA topic and word distributions over time. Both these models do not allow tracking specific topics of interest to the user and only model the evolution of topics that are more dominant in the data and tend to ignore rarer topics. Hence, topic evolution models such as DTM and ToT are not effective in this setting. Hall et al. [8] analyze the history of ideas in NLP conferences using topic modeling. They use LDA to model topic evolution across years and use hand-selected seed words to track evolution of specific topics. They also note that the temporal variants of LDA (DTM and ToT) impose constraints on the time periods, rendering them inflexible and unsuitable for modeling documents that can change dramatically from one time period to another. In online courses, each course iteration attracts an entirely new cohort of students and the content in the forums can potentially vary significantly according to their interests and backgrounds. Hence, seeding topic models with words

is a simple, yet effective means to track evolution of specific topics of interest. In our work, we leverage a seeded variant of LDA, Seeded LDA [9] to track specific topics of conversation in the forums across iterations. Seeded LDA guides topic discovery to learn specific topics of interest by allowing the user to input a set of seed words that are representative of the underlying topics in the corpus. Seeded LDA uses these seed words to improve topic-word distribution by inducing topics to obtain a high probability mass for the given seed words. Similarly, it also improves the document-topic distribution in online courses by biasing documents to select topics related to the seed words. The seed set need not be exhaustive as the model gathers related words based on co-occurrence of other words with the specified seed words in the documents. We refer the reader to [9] for more details.

3 Data

We analyze data from two popular long running Coursera MOOCs: a) a business course, and b) a computer science course. We refer to these courses as BUSINESS course and CS course, respectively. Both courses are active courses attracting thousands of students every iteration. We analyze 34 iterations of the BUSINESS course, each iteration spanning 6 weeks. The BUSINESS course has on an average is greater than $50,000$ users and $5,000$ posts per iteration. The highest number of students registered is approximately $100,000$ students and the corresponding discussion forum has $10,000$ posts. Hence, in total, we analyze approximately $200,000$ posts across all iterations of BUSINESS course. The CS course spans 8 weeks for the first three iterations. From the fourth iteration, the course splits into two parts spanning 6 weeks each, which we refer to as CS-1 and CS-2. In all, we analyze data from 15 course offerings, 3 iterations of the original course, and 6 iterations each of CS-1 and CS-2. The CS course has on an average is greater than $100,000$ users and $13,000$ posts per iteration. The highest number of students registered is approximately $250,000$ students and the corresponding discussion forum has $47,000$ posts. Hence, in total, we analyze approximately $110,000$ posts across all iterations of CS course.

4 Topic Discovery in Online Courses

In this section, we build the seeded topic models for discovering topics in forum posts. Due to the absence of labeled data, the seed words are hand-selected. We construct *four* seeded LDA models using four different combinations of seeded topics. Ramesh et al. [11,10] give the seed words for the different coarse and fine-grained topics. For each seeded LDA model, we include an additional k unseeded topics in the seeded LDA models to capture topics that do not fall under the seeded topic categories. After experimenting with different values of k, we choose $k = 5$. Hence, the number of topics for each model is the sum of number of seeded topics and 5 unseeded topics. We train all our seeded LDA models for 1000 iterations. We use α = 0.0001 and β = 0.0001 to create a sparse topic distribution so that fewer topics with high values emerge.

4.1 Primary Purpose of Forums

In the first categorization, we classify the posts into three categories: i) social, ii) issues, and iii) course content topics. These three topic categories reflect the three primary purposes in which the forums are utilized in online courses. Social posts capture the

social aspect of forums, where students can e-socialize with fellow classmates. These posts usually fall into one of the following subcategories: a) student introductions, and b) formation of study groups. Issue posts are posts that intend to bring issues in the course to the attention of the instructor and fellow classmates and/or ask for their help in solving them.

4.2 Course Elements

Unlike most classroom courses, online courses attract a diverse set of students with varied interests and expectations from the course. Anderson et al. [2] classify students according to their interaction on the MOOC. Of these, three most common types of students include: 1) *viewers*: students interested in *viewing* video lectures, 2) *solvers*: students interested in *solving* assignments, and 3) students interesting in obtaining a certificate. These three types of students map to the three corresponding course elements: i) lectures, ii) quizzes/assignments, and iii) certificate. In the second categorization, we identify posts corresponding to these three important course elements. Analyzing references to these elements in the forums helps us understand the different types of students in the course and which course elements to focus on for future iterations.

4.3 Fine-grained Analysis of Issue Posts

In the third categorization, we further drill down on issue posts to understand how they are distributed across course elements. For this, we combine the two seeded LDA models to categorize issue posts further across the three course elements. We label posts for which *issue* topic has the highest value in the document-topic distribution as *issue* posts and further categorize these posts across the three course elements.

We identify fine-grained sub-topics for *lecture*, *video* and *subtitles*, and *quiz*, *submission*, *grading*, and *deadline*, and categorize the logistic issue posts into these fine-grained topics. Similarly, in the fourth categorization, we combine the seeded LDA models for *issue* with fine-grained seeded LDA models to understand how *issue* posts are distributed across fine-grained lecture and quiz topics.

5 Topic Trends in Online Courses

In this section, we present an in-depth analysis of topic evolution across iterations of the BUSINESS course and the CS course. For each course, we conduct experiments to answer the following questions:

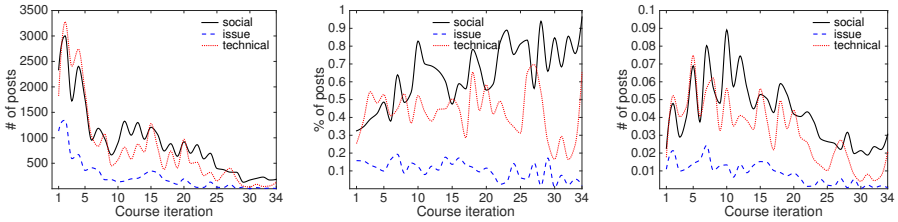
1. How are posts distributed across topics constituting the three primary purposes of forums: a) social, b) issues, and c) technical topics, and how is that evolving with time?
2. Next, we answer the question of which course elements are most popular in the course and how is the emphasis on each of them changing with time?
3. Finally, we drill down deeper on issue posts and analyze which topics constitute the focus of issue posts and how are they changing as iterations unfold?

We run seeded LDA models described in Sections 4.1, 4.2, and 4.3 across iterations of the BUSINESS course and the CS course to answer the above three questions.

5.1 BUSINESS course

Here, we present topic evolution analysis of posts in the BUSINESS course.

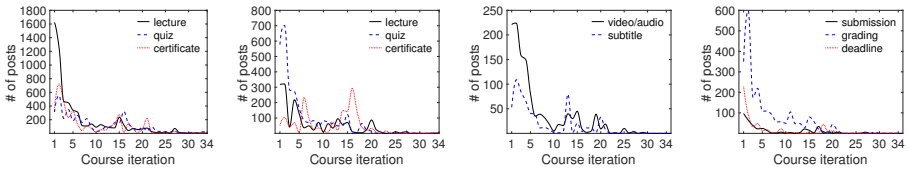
Primary Purpose of Forums In our first set of experiments, we study the evolution of social, issue, and technical topics across iterations. For each iteration of the course, we add the topic distribution values for each topic across all the posts to get the total number of posts in each topic category. We then plot the number of posts in each topic category across iterations. Figure 1(a) gives the number of *social*, *issue*, and *technical* posts across iterations. In the BUSINESS course, we observe that social posts contribute to a significant number of posts in the forum, emphasizing the importance of forums as a socializing platform. This is closely followed by technical posts. Issue posts are fewer in number in comparison to social and technical posts and decline to negligible numbers in the later iterations. Social and issue posts also decline over time, but always remain higher than issue posts in the later iterations.



(a) # social, issue, and technical posts (b) % of social, issue, and technical posts (c) # social, issue, and technical posts per student

Fig. 1. BUSINESS course: evolution of social, issue, and technical posts across iterations

Analyzing the percentage of social, issue, and technical posts in the total number of posts in each iteration (Figure 1(b)), we observe that social and technical topics together constitute around 80% of posts. The percentage of social posts continuously increases, emphasizing the importance of the social aspect in learning. The percentage of technical discussions follows a steady path of evolution across iterations and they constitute a significant percentage of forum discussions even in the later iterations, which helps us to understand that there is a significant amount of interest in the technical course content. Issues contribute to less than 20% of posts in the early iterations, declining steadily, dropping to less than 10% after 30 iterations. Analyzing the number of social and issue posts per student (Figure 1(c)), we observe that an increase across all three categories steadily in the initial iterations followed by a steady decline, indicating that fewer students tend to post in the forums as the course stabilizes.



(a) # lecture, quiz, and certificate posts (b) # lecture, quiz, and certificate *issue* posts (c) # lecture issue posts (d) # quiz issue posts

Fig. 2. BUSINESS course: distribution of posts and *issue* posts across three course elements: lecture, quiz, and certificate, and lecture and quiz sub-topics

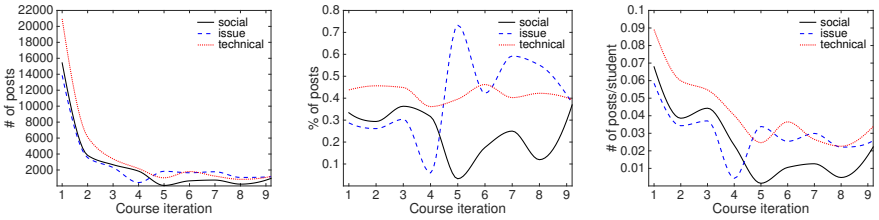
Course Elements In our second set of experiments, we analyze the emphasis on different course elements across iterations. This analysis is helpful to understand what are the most sought after course elements in this course. Figure 2(a) gives the number of posts in the three course elements across iterations. We observe that lectures emerge as the most dominant course element in BUSINESS course across all iterations, followed by quiz and certificate, in that order. We observe that certificates are a popular topic category in BUSINESS course, consistently attracting posts throughout all iterations.

Fine-grained Analysis of Issue Posts In our third set of experiments, we analyze how issue posts are distributed across the course elements. Figure 2(b) gives the distribution of issues across the course elements. It is interesting to note that while lectures are the most discussed course element, most issues are reported on quizzes in the initial iterations.

While we observe a consistent interest in certificates across all iterations, there are two periods which show an increased incidence of certificate issues: around 6^{th} iteration and 16^{th} iteration. Analyzing the certificate issue posts in these iterations, we find that there was a delay in dispatching certificates in both these periods, causing a flurry of certificate issue posts.

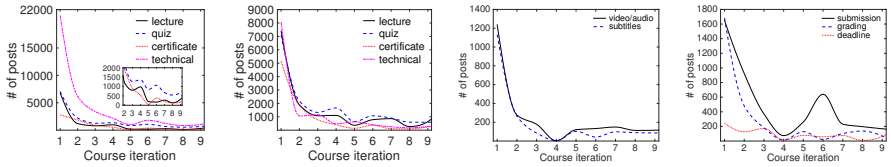
Next, we perform a finer-grained analysis of issue posts across fine-grained lecture and quiz topics. Figure 2(c) gives the distribution of issues across lecture sub-topics: video/audio and subtitles. We notice that video/audio issues are more prominent in the earlier iterations. Both video/audio and subtitle issues decline and contribute almost equally to lecture issues in the middle iterations before declining to negligible number of posts in the later iterations.

Figure 2(d) gives the distribution of issues across quiz sub-topics. We observe that a major proportion of quiz issues fall under grading, with submission and deadline hardly contributing to quiz-related issues. Grading issues follow a steep decline from the third iteration after peer grading was replaced with automatic grading in the course, indicating a preference among students for the latter. Often instructors make modifications to the course responding to feedback from students. Our analysis not only helps them identify the issues but also provides them with a simple and effective tool to evaluate the success of their improvements.



(a) # social, issue, and technical posts (b) % of social, issue, and technical posts (c) # social, issue, and technical posts per student

Fig. 3. CS course: evolution of social, issue, and technical posts across iterations



(a) # lecture, quiz, certificate, and technical posts

(b) # lecture, quiz, certificate, and technical *issue* posts

(c) # lecture related issue posts

(d) # quiz related issue posts

Fig. 4. CS course: distribution of *issue* posts across fine-grained topics

5.2 CS course

Here, we present topic evolution results for the CS course. In all, we analyze data from 15 course offerings, 3 iterations of the original course and 6 iterations each of CS-1 and CS-2, respectively. We coalesce CS-1 and CS-2 for each iteration and treat that as a single course, giving us *nine* iterations of the CS course.

Primary Purpose of Forums Figure 3(a) gives the number of posts in the social, issue, and technical topics across iterations in the CS course. For the technical topic, we add the values in the document-topic multinomial distribution given by seeded LDA across CS technical seeded topics and present their combined evolution over time. We notice a different trend in the CS course when compared to the BUSINESS course. While social posts dominate the forums in the initial iterations, they slowly decline from the 4^{th} iteration. Technical and issue posts primarily dominate the forums from the 5^{th} iteration, with issue posts being more predominant when compared to technical posts.

Figure 3(b) gives the percentage of social and issue posts in the forums. We again observe that as iterations unfold, the percentage of technical posts remains the same, while there is a marked increase in issue posts. Percentage of issue posts peaks in the 5^{th} iteration and declines thereafter, but still remains higher than social and technical posts. It is also interesting to note that this increase happens immediately after the course splits (4^{th} iteration). Intuitively, as courses stabilize, it is expected that issues reported in the previous iterations are fixed causing issue posts to decrease over time. But in the CS course, we observe the opposite, which calls for a detailed analysis of why issue posts exhibit an increasing trend and what kind of issues are being reported by students. Another interesting trend to note is that in Figure 3(c), we observe that higher percentage of issue posts come from a fewer number of students when compared to the technical posts.

Issues reported in the CS course vary significantly in comparison to issues in BUSINESS course. Computer science courses often have software installation prerequisites that could potentially trigger a large number of posts around errors in installing/compiling software. Unlike logistic issues, these issues are inherently different in nature and in most cases cannot be easily fixed by the instructor, especially in an online setting.

Course Elements In our next set of experiments, we analyze the evolution of topics corresponding to the three course elements. We add the *technical* topic to this classification for readability, as we will be drilling deeper into the technical issues along with issues in course elements in Section 5.2.

Figure 4(a) gives the evolution of course elements as iterations unfold. Technical topic attracts the most number of posts in the CS course across all iterations. Concentrating on the zoomed in portion of Figure 4(a), we observe that among the three course elements, quizzes emerge as the most dominating course element in the CS course. While lectures are the most sought after course element in the BUSINESS course, they rank second in popularity in the CS course, and this is followed by certificate. This analysis helps instructors prioritize and focus limited resources on course elements that students care about the most.

Fine-grained Analysis of Issue Posts Next, we investigate how issues are distributed across the course elements. We add the *technical* topic category to the list of course elements to model the evolution of technical issue posts. Figure 4(b) gives the evolution of issues across lecture, quiz, certificate, and technical topics. We find that technical issues dominate issue posts across all iterations, followed by quiz related issues. At iteration 5, where we observe an overall increase in issues in Figure 3, we observe a similar spike in the quiz and technical issue posts in Figure 4(b) as well. A plausible reason for this increase is that when the course splits into two courses in the 4^{th} iteration, more programming assignments were added, which led to more technical and quiz issues to be reported. Lecture and certificate topics hardly contribute to the issue posts and decline to small numbers as iterations progress.

We further break down issues across lecture and quiz sub-topics in Figure 4. Figure 4(c) gives the distribution of issue posts across lecture sub-topics: video/audio and subtitles. As we observed in Figure 4(b), there are only a few lecture issue posts in each iteration and this reflects in the finer analysis as well. Between the lecture subtopics, video/audio is the most contributing sub-category. Performing a similar analysis on quiz sub-topics (Figure 4(d)), we find that most of the quiz issue posts fall under the submission category, which is followed by grading, and deadlines. While grading consistently remains a contributing issue category across both the courses, we note that the structure of CS course requires submitting computer programs in an online platform inciting a significant number of issue posts in the submission category. Isolating logistic issues in the CS course and comparing this to Figure 1(a), we find they follow a similar pattern to the BUSINESS course, declining over time. This supports our hypothesis that as courses stabilize lesser logistic issues surface and hence they are reported less in the forums.

6 Conclusion

In this work, we presented a detailed temporal analysis of two long-running MOOCs from different disciplines and identified the similarities and differences between them. Our methodology and analysis is helpful in determining the stability of these courses and identifying opportunities for improvement. The weakly supervised nature of our approach using a common set of seed words is beneficial in extending our models with minimal effort to analyze newer courses. There are several exciting future research directions. The temporal analysis can potentially be integrated with an automatic feedback

mechanism to actively monitor student feedback. This can be especially helpful when large course changes (such as splitting the course, or changing the grading methodology) are deployed. This feedback mechanism can help instructors get notified of abrupt changes in the forums and allow them to address concerns promptly, thus helping in improving students' interaction experience.

References

1. Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. In *Proceedings of the International Conference on Educational Data Mining (EDM)*.
2. Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with Massive Online Courses. In *Proceedings of the International Conference on World Wide Web (WWW)*.
3. David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.
4. Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. 2014. Predicting Instructor's Intervention in MOOC forums. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
5. Yi Cui and Alyssa Friend Wise. 2015. Identifying Content-Related Threads in MOOC Discussion Forums. In *Proceedings of the ACM Conference on Learning @ Scale (L@S)*.
6. Aysu Ezen-Can, Kristy Elizabeth Boyer, Shaun Kellogg, and Sherry Booth. 2015. Unsupervised Modeling for Understanding MOOC Discussion Forums: A Learning Analytics Approach. In *Proceedings of the International Conference on Learning Analytics And Knowledge (LAK)*.
7. Andr Gohr, Alexander Hinneburg, Ren Schult, and Myra Spiliopoulou. 2009. Topic Evolution in a Stream of Documents. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*.
8. David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
9. Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.
10. Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2014. Understanding MOOC Discussion Forums using Seeded LDA. In *9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*. ACL.
11. Arti Ramesh, Shachi Kumar, James Foulds, and Lise Getoor. 2015. Weakly Supervised Models of Aspect-Sentiment for Online Course Discussion Forums. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
12. Xuerui Wang and Andrew McCallum. 2006. Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
13. Jian-Syuan Wong, Bart Pursel, Anna Divinsky, and Bernard J. Jansen. 2015. An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. In *Proceedings of the Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP)*.