Understanding MOOC Discussion Forums using Seeded LDA

[1]Arti Ramesh, [1]Dan Goldwasser, [1]Bert Huang, [1]Hal Daumé III, [2]Lise Getoor
[1]University of Maryland, College Park [2]University of California, Santa Cruz
{artir, bert, hal}@cs.umd.edu, goldwas1@umiacs.umd.edu, getoor@soe.ucsc.edu

Abstract

Discussion forums serve as a platform for student discussions in massive open online courses (MOOCs). Analyzing content in these forums can uncover useful information for improving student retention and help in initiating instructor intervention. In this work, we explore the use of topic models, particularly *seeded topic models* toward this goal. We demonstrate that features derived from topic analysis help in predicting *student survival*.

1 Introduction

This paper highlights the importance of understanding MOOC discussion forum content, and shows that capturing discussion forum content can help uncover students' intentions and motivation and provide useful information in predicting course completion.

MOOC discussion forums provide a platform for exchange of ideas, course administration and logistics questions, reporting errors in lectures, and discussion about course material. Unlike classroom settings, where there is face-to-face interaction between the instructor and the students and among the students, MOOC forums are the primary means of interaction in MOOCs. However, due to the large number of students and the large volume of posts generated by them, MOOC forums are not monitored completely. Forums can include student posts expressing difficulties in course-work, grading errors, dissatisfaction in the course, which are possible precursors to students dropping out.

Previous work analyzing discussion forum content tried manually labeling posts by categories of interest (Stump et al., 2013). Unfortunately, the effort involved in manually annotating the large amounts of posts prevents using such solutions on a large scale. Instead, we suggest using natural language processing tools for identifying relevant aspects of forum content automatically. Specifically, we explore *SeededLDA* (Jagarlamudi et al., 2012), a recent extension of topic models which can utilize a lexical seed set to bias the topics according to relevant domain knowledge.

Exploring data from three MOOCs, we find that forum posts usually belong to these three categories—a) course content, which include discussions about course material (COURSE), b) meta-level discussions about the course, including feedback and course logistics (LOGISTICS), and c) other general discussions, which include student introductions, discussions about online courses (GENERAL). In order to capture these categories automatically we provide seed words for each category. For example, we extract seed words for the COURSE topic from each course's syllabus. In addition to the automatic topic assignment, we capture the sentiment polarity using *Opinionfinder* (Wilson et al., 2005). We use features derived from topic assignments and sentiment to predict student course completion *(student survival)*. We measure course completion by examining if the student attempted the final exam/ last few assignments in the course. We follow the observation that LOGISTICS posts contain feedback about the course. Finding high-confidence LOGISTICS posts can give a better understanding of student opinion about the course. Similarly, posting in COURSE topic and receiving good feedback (i.e., votes) is an indicator of student success and might contribute to survival. We show that modeling these intuitions using topic assignments together with sentiment scores, helps in predicting student survival. In addition, we examine the topic assignment and sentiment patterns of some users and show that topic assignments help in understanding student concerns better.

2 Modeling Student Survival

Our work builds on work by Ramesh et al. (2013) and (2014) on modeling student survival using Probabilistic Soft Logic (PSL). The authors included behavioral features, such as lecture views, posting/voting/viewing discussion forum content, linguistic features, such as sentiment and subjectivity of posts, and social interaction features derived from forum interaction. The authors looked at indication of sentiment without modeling the context in which the sentiment was expressed: positive sentiment implying survival and negative sentiment implying drop-out. In this work, we tackle this problem by adding topics, enabling reasoning about specific types of posts. While sentiment of posts can indicate general dissatisfaction, we expect this to be more pronounced in LOGISTICS posts as posts in this category correspond to issues and feedback about the course. In contrast, sentiment in posts about course material may signal a particular topic of discussion in a course and may not indicate attitude of the student toward the course. In Section 4.3, we show some examples of course-related posts and their sentiment, and we illustrate that they are not suggestive of student survival. For example, in *Women and the Civil Rights Movement* course, the post—*"I think our values are shaped by past generations in our family as well, sometimes negatively."*—indicates an attitude towards an issue discussed as part of the course. Hence, identifying posts that fall under LOGISTICS can improve the value of sentiment in posts. In Section 3, we show how these are translated into rules in our model.

2.1 Probabilistic Soft Logic

We briefly overview the some technical details behind Probabilistic Soft Logic (PSL). For brevity, we omit many specifics, and we refer the reader to (Broecheler et al., 2010; Bach et al., 2013) for more details. PSL is a framework for collective, probabilistic reasoning in relational domains. Like other statistical relational learning methods (Getoor and Taskar, 2007), PSL uses weighted rules to model dependencies in a domain. However, one distinguishing aspect is that PSL uses continuous variables to represent truth values, relaxing Boolean truth values to the interval [0,1].

Table 1 lists some PSL rules from our model. The predicate *posts* captures the relationship between a post and the user who posted it. Predicate *polarity(P)* represents sentiment via its truth value in $[0, 1]$, where 1.0 signifies positive sentiment, and 0.0 signifies negative sentiment. *upvote(P)* is 1.0 if the post has positive feedback and 0.0 if the post had negative or no feedback. *U* and *P* refer to *user* and *post* respectively. These features can be combined to produce rules in Table 1. For example, the first rule captures the idea that posts with positive sentiment imply student survival.

- $posts(U, P) \wedge polarity(P) \rightarrow survival(U)$
- $posts(U, P) \wedge \neg polarity(P) \rightarrow \neg survival(U)$
- $posts(U, P) \wedge upvote(P) \rightarrow survival(U)$

Table 1: Example rules in PSL

3 Enhancing Student Survival Models with Topic Modeling

Discussion forums in online courses are organized into threads to facilitate grouping of posts into topics. For example, a thread titled *errata, grading issues* is likely a place for discussing course logistics and a thread titled *week 1, lecture 1* is likely a place for discussing course content. But a more precise examination of such threads reveals that these heuristics do not always hold. We have observed that *course content* threads often house *logistic content* and vice-versa. This demands the necessity of using computational linguistics methods to classify the content in discussion forums.

In this work, we—1) use topic models to map posts to topics in an unsupervised way, and 2) employ background knowledge from the course syllabus and manual inspection of discussion forum posts to seed topic models to get better separated topics. We use data from three Coursera MOOCs: *Surviving Disruptive Technologies*, *Women and the Civil Rights Movement*, and *Gene and the Human Condition* for our analysis. In discussion below, we refer to these courses as DISRTECH, WOMEN, and GENE, respectively.

3.1 Latent Dirichlet Allocation

Table 2 gives the topics given by *latent Dirichlet allocation* (LDA) on discussion forum posts. The words that are likely to fall under LOGISTICS are underlined in the table. It can be observed that these words are spread across more than one topic. Since we are especially interested in posts that are on LOGISTICS, we use *SeededLDA* (Jagarlamudi et al., 2012), which allows one to specify *seed* words that can influence the discovered topics toward our desired three categories.

topic 1:	kodak, management, great, innovation, post, agree, film, understand, something, problem, businesses, changes, needs
topic 2:	good, change, publishing, brand, companies, publishers, history, marketing, traditional, believe, authors
topic 3:	think, work, technologies, newspaper, content, paper, model, business, disruptive, information, survive, print, media, course, assignment
topic 4:	digital, kodak, company, camera, market, quality, phone, development, future, failed, high, right, old,
topic 5:	amazon, books, netflix, blockbuster, stores, online, experience, products, apple, nook, strategy, video, service
topic 6:	time, grading, different, class, course, major, focus, product, like, years
topic 7:	companies, interesting, class, thanks, going, printing, far, wonder, article, sure

Table 2: Topics identified by LDA

topic 1:	thank, professor, lectures, assignments, concept, love, thanks, learned, enjoyed, forums, subject, question, hard, time, grading, peer, lower, low
topic 2:	learning, education, moocs, courses, students, online, university, classroom, teaching, coursera

Table 3: Seed words in LOGISTICS and GENERAL for DISR-TECH, WOMEN and GENE courses

topic 3a:	disruptive, technology, innovation, survival, digital, disruption, survivor
topic 3b:	women, civil, rights, movement, american, black, struggle, political, protests, organizations, events, historians, african, status, citizenship
topic 3c:	genomics, genome, egg, living, processes, ancestors, genes, nature, epigenitics, behavior, genetic, engineering, biotechnology

Table 4: Seed words for COURSE topic for DISR-TECH, WOMEN and GENE courses

topic 1:	time, thanks, one, low, hard, question, course, love, professor, lectures, lower, another, concept, agree, peer, point, never
topic 2:	online, education, coursera, students, university, courses, classroom, moocs, teaching, video
topic 3:	digital, survival, management, disruption, technology, development, market, business, innovation
topic 4:	publishing, publisher, traditional, companies, money, history, brand
topic 5:	companies, social, internet, work, example
topic 6:	business, company, products, services, post, consumer, market, phone, changes, apple
topic 7:	amazon, book, nook, readers, strategy, print, noble, barnes

Table 5: Topics identified by SeededLDA for DISR-TECH

topic 1:	time, thanks, one, hard, question, course, love, professor, lectures, forums, help, essays, problem, thread, concept, subject
topic 2:	online, education, coursera, students, university, courses, classroom, moocs, teaching, video, work, english, interested, everyone
topic 3:	women, rights, black, civil, movement, african, struggle, social, citizenship, community, lynching, class, freedom, racial, segregation
topic 4:	violence, public, people, one, justice, school,s state, vote, make, system, laws
topic 5:	idea, believe, women, world, today, family, group, rights
topic 6:	one, years, family, school, history, person, men, children, king, church, mother, story, young
topic 7:	lynching, books, mississippi, march, media, youtube, death, google, woman, watch, mrs, south, article, film

Table 6: Topics identified by SeededLDA for WOMEN

topic 1:	time, thanks, one, answer, hard, question, course, love, professor, lectures, brian, lever, another, concept, agree, peer, material, interesting
topic 2:	online, education, coursera, students, university, courses, classroom, moocs, teaching, video, knowledge, school
topic 3:	genes, genome, nature, dna, gene, living, behavior, chromosomes, mutation, processes
topic 4:	genetic, biotechnology, engineering, cancer, science, research, function, rna
topic 5:	reproduce, animals, vitamin, correct, term, summary, read, steps
topic 6:	food, body, cells, alleles blood, less, area, present, gmo, crops, population, stop
topic 7:	something, group, dna, certain, type, early, large, cause, less, cells

Table 7: Topics identified by SeededLDA for GENE

3.2 Seeded LDA

We experiment by providing seed words for topics that fall into the three categories. The seed words for the three courses are listed in tables 3 and 4. The seed words for LOGISTICS and GENERAL are common across all the three courses. The seed words for the COURSE topic are chosen from the course-syllabus of the courses. This construction of seed words enables the model to be applied to new courses easily. Topics *3a*, *3b*, and *3c* denote the course specific seed words for DISR-TECH, WOMEN, and GENE courses respectively. Since the syllabus is only an outline of the class, it does not contain all the terms that will be used in class discussions. To capture other finer course content discussions as separate topics, we include k more topics when we run the SeededLDA. We notice that not including more topics here, only including the seeded topics (i.e., run SeededLDA with exactly three topics) results in some words from course content discussions, which were not specified in the course-seed words, appearing in the LOGISTICS or GENERAL topics. Thus, the k extra topics help represent COURSE topics that do not directly correspond to the course seeds. Note that these *extra* topics are not seeded. We experimented with different values of k on our experiments and found by manual inspection that the topic-terms produced by our model were well separated for $k = 3$. Thus, we run *SeededLDA* with 7 total topics. Tables 5, 6, and 7 give the topics identified for DISR-TECH, WOMEN and GENE by *SeededLDA*. The topic assignments so obtained are used as input features to the PSL model—the predicate for the first topic is LOGISTICS, the second one is GENERAL and the rest are summed up to get the topic assignment for COURSE.

3.3 Using topic assignments in PSL

We construct two models—a) DIRECT model, including all features except features from topic

survival = 0.0	polarity = 0.25	logistics = 0.657
		general = 0.028
		course = 0.314
			JSTOR allowed 3 items (texts/writings) on my 'shelf' for 14 days. But, I read the items and wish to return them, but cannot, until 14 days has expired. It is difficult then, to do the extra readings in the "Exploring Further" section of Week 1 reading list in a timely manner. Does anyone have any ideas for surmounting this issue?

survival = 0.0	polarity = 0.0	logistics = 0.643
		general = 0.071
		course = 0.285
			There are some mistakes on quiz 2. Questions 3, 5, and 15 mark you wrong for answers that are correct.

survival = 0.0	polarity = 0.25	logistics = 0.652
		general = 0.043
		course = 0.304
			I see week 5 quiz is due April 1(by midnight 3/31/13).I am concerned about this due date being on Easter, some of us will be traveling, such as myself. Can the due date be later in the week? Thank you

Table 8: Logistics posts containing negative sentiment for dropped-out students

survival = 1.0	polarity = 0.0	logistics = 0.67
		general = 0.067
		course = 0.267
			I was just looking at the topics for the second essay assignments. The thing is I dont see what the question choices are. I have the option of Weeks and I have no idea what that even means. Can someone help me out here and tell me what the questions for the second essay assignment are I think my computer isnt allowing me to see the whole assignment! Someone please help me out and let me know that the options are.

survival = 1.0	polarity = 0.25	logistics = 0.769
		general = 0.051
		course = 0.179
			I'd appreciate someone looks into the following: Lecture slides for the videos (week 5) don't open (at all) (irrespective of the used browser). Some required reading material for week 5 won't open either (error message). I also have a sense that there should be more material posted for the week (optional readings, more videos, etc). Thanks. — I am not seeing a quiz posted for Week 5.

survival = 1.0	polarity = 0.78	logistics = 0.67
		general = 0.067
		course = 0.267
			Hopefully the Terrell reading and the Lecture PowerPoints now open for you. Thanks for reporting this.

Table 9: Example of change in sentiment in a course logistic thread

survival = 1.0	polarity = 0.25	logistics = 0.372
		general = 0.163
		course = 0.465
			I've got very interested in the dynamic of segregation in terms of space and body pointed by Professor Brown and found a document written by GerShun Avilez called "Housing the Black Body: Value, Domestic Space,and Segregation Narratives".

survival = 1.0	polarity = 0.9	logistics = 0.202
		general = 0.025
		course = 0.772
			I think that you hit it on the head, the whole idea of Emancipation came as a result not so much of rights but of the need to get the Transcontinental Railroad through the mid-west and the north did not want the wealth of the southern slave owners to overshadow the available shares. There are many brilliant people "good will hunting", and their brilliance either dies with them or dies while they are alive due to intolerance. Many things have happened in my life to cause me to be tolerant to others and see what their debate is, Many very evil social ills and stereotypes are a result of ignorance. It would be awesome if the brilliant minds could all come together for reform and change.

survival = 1.0	polarity = 0.167	logistics = 0.052
		general = 0.104
		course = 0.844
			I think our values are shaped by past generations in our family as well – sometimes negatively. In Bliss, Michigan where I come from, 5 families settled when the government kicked out the residents – Ottowa Tribe Native Americans. I am descended from the 5 families. All of the cultural influences in Bliss were white Christian – the Native American population had never been welcomed back or invited to stay as they had in Cross Village just down the beach. My family moved to the city for 4 years during my childhood, and I had African American, Asian, and Hispanic classmates and friends. When we moved back to the country I was confronted with the racism and generational wrong-doings of my ancestors. At the tender age of 10 my awareness had been raised! Was I ever pissed off when the full awareness of the situation hit me! I still am.

Table 10: Posts talking about COURSE content

DIRECT	DIRECT+TOPIC
$posts(U,P) \wedge polarity(P) \rightarrow survival(U)$	$posts(U,P) \wedge topic(P, \text{LOGISTICS}) \wedge \neg polarity(P) \rightarrow survival(U)$
$posts(U,P) \wedge \neg polarity(P) \rightarrow \neg survival(U)$	$posts(U,P) \wedge topic(P, \text{LOGISTICS}) \wedge \neg polarity(P) \rightarrow survival(U)$
$posts(U,P) \rightarrow survival(U)$	$posts(U,P) \wedge topic(P, \text{GENERAL}) \rightarrow \neg survival(U)$
$posts(U,P) \wedge upvote(P) \rightarrow survival(U)$	$posts(U,P) \wedge topic(P, \text{COURSE}) \wedge upvote(P) \rightarrow survival(U)$
	$posts(U_1,P) \wedge posts(U_2,P) \wedge topic(P, \text{COURSE}) \wedge survival(U_1) \rightarrow survival(U_2)$

Table 11: Rules modified to include topic features

modeling, and b) DIRECT+TOPIC model, including the topic assignments as features in the model. Our DIRECT model is borrowed from Ramesh (2014). We refer the reader to (Ramesh et al., 2013) and (Ramesh et al., 2014) for a complete list of features and rules in this model.

Table 11 contains examples of rules in the DIRECT model and the corresponding rules including topic assignments in DIRECT+TOPIC model. The first and second rules containing polarity are changed to include LOGISTICS topic feature, following our observation that polarity matters in *meta-course* posts. While the DIRECT model regards posting in forums as an indication of survival, in the DIRECT+TOPIC model, this rule is changed to capture that students that post a lot of *general* stuff *only* on the forums do not necessarily participate in course-related discussions. The fourth rule containing *upvote* predicate, which signifies posts that received positive feedback in the form of votes, is changed to include the topic-feature COURSE. This captures the significance of posting *course-related* content that gets positive feedback as opposed to *logistics* or *general* content in the forums. This rule helps us discern posts in general/logistic category that can get a lot of positive votes (*upvote*), but do not necessarily indicate student survival. For example, some introduction threads have a lot of positive votes, but do not necessarily signify student survival.

4 Empirical Evaluation

We conducted experiments to answer the following question—how much do the topic assignments from *SeededLDA* help in predicting student survival? We also perform a qualitative analysis of topic assignments, the sentiment of posts, and their correspondence with student survival.

COURSE	MODEL	AUC-PR POS.	AUC-PR NEG.	AUC-ROC
DISR-TECH	DIRECT	0.764	0.628	0.688
	DIRECT+TOPIC	**0.794**	0.638	**0.708**
WOMEN	DIRECT	0.654	0.899	0.820
	DIRECT+TOPIC	**0.674**	0.900	**0.834**
GENE	DIRECT	0.874	0.780	0.860
	DIRECT+TOPIC	**0.894**	**0.791**	**0.873**

Table 12: Performance of DIRECT and DIRECT+TOPIC models in predicting student survival. Statistically significant scores typed in bold.

4.1 Datasets and Experimental Setup

We evaluate our models on three Coursera MOOCs: DISR-TECH, WOMEN-CIVIL, and GENE, respectively. Our data consists of anonymized student records, grades, and online behavior recorded during the seven week duration of each course. We label students as *survival = 1.0* if they take the final exam/quiz and *survival = 0.0* otherwise. In our experiments, we only consider students that completed at least *one* quiz/assignment. We evaluate our models using area under precision-recall curve for positive and negative *survival* labels and area under ROC curve. We use ten-fold cross-validation on each of the courses, leaving out 10% of users for testing and revealing the rest of the users for training the model weights. We evaluate statistical significance using a paired t-test with a rejection threshold of *0.05*.

4.2 Survival Prediction using topic features

Table 12 shows the prediction performance of the DIRECT and DIRECT+TOPIC model. The inclusion of the topic-features improves student survival prediction in all the three courses.

4.3 Discussion topic analysis using topic features

Table 8 shows some posts by users that did not survive the class. All these posts have negative sentiment scores by *Opinionfinder* and belong to LOGISTICS. Also, in the forum, all these posts were not answered. This suggests that students might drop out if their *course-logistics* questions are not answered. Table 9 gives examples of student posts that also have a negative sentiment. But the sentiment of the thread changes when the issue is resolved (last row in the table). We observe that these two students survive the course and a timely answer to their posts might have been a reason influencing these students to complete the course.

Tables 8 and 9 show how student survival may depend on forum interaction and responses they receive. Our approach can help discover potential points of contention in the forums, identifying potential drop outs that can be avoided by intervention.

Table 10 shows posts flagged as COURSE by the *SeededLDA*. The polarity scores in the COURSE posts indicate opinions and attitude toward course specific material. For example, post #3 in Table 10 indicates opinion towards human rights. While the post's polarity is negative, it is clear that this polarity value is not directed at the course and should not be used to predict student survival. In fact, all these users survive the course. We find that participation in course related discussion is a sign of survival. These examples demonstrate that analysis on COURSE posts can mislead survival and justify our using topic predictions to focus sentiment analysis on LOGISTICS posts.

5 Discussion

In this paper, we have taken a step toward understanding discussion content in massive open online courses. Our topic analysis is coarse-grained, grouping posts into three categories. In our analysis, all the meta-content—course logistics and course feedback—were grouped under the same topic category. Instead, a finer-grained topic model could be seeded with different components of meta-content as separate topics. The same applies for course-related posts too, where a finer-grained analysis could help identify difficult topics that may cause student frustration and dropout.

Acknowledgements We thank the instructors for letting us use data from their course. This work is supported by National Science Foundation under Grant No. CCF0937094. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. 2013. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence*.

Matthias Broecheler, Lilyana Mihalkova, and Lise Getoor. 2010. Probabilistic similarity logic. In *Uncertainty in Artificial Intelligence (UAI)*.

Lise Getoor and Ben Taskar. 2007. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 204–213, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angelika Kimmig, Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2013. Modeling learner engagement in MOOCs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2014. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Glenda S. Stump, Jennifer DeBoer, Jonathan Whittinghill, and Lori Breslow. 2013. Development of a framework to classify mooc discussion forum posts: Methodology and challenges. In *NIPS Workshop on Data Driven Education*.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*.