Understanding MOOC Discussion Forums using Seeded LDA

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, Lise Getoor
University of Maryland, College Park  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].

Moreover, one distinguishing aspect is that 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 \( \text{posts} \) captures the relationship between a post and the user who posted it. Predicate \( \text{polarity}(P) \) represents sentiment via its truth value in \([0, 1]\), where 1.0 signifies positive sentiment, and 0.0 signifies negative sentiment. \( \text{upvote}(P) \) 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.

| Rule | Description |
|------|-------------|
| \( \text{posts}(U, P) \land \text{polarity}(P) \rightarrow \text{survival}(U) \) | posts imply survival |
| \( \text{posts}(U, P) \land \neg \text{polarity}(P) \rightarrow \neg \text{survival}(U) \) | negative polarity implies drop-out |
| \( \text{posts}(U, P) \land \text{upvote}(P) \rightarrow \text{survival}(U) \) | upvoted posts imply survival |

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 DISR-TECH, 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.
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
The following our observation that polarity matters in changed to include LOGISTICS. The first and second rules containing polarity are DIRECTing topic assignments in 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.

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. Opinion-Finder: A system for subjectivity analysis. In Proceedings of HLT/EMNLP on Interactive Demonstrations.