Predicting Instructor's Intervention in MOOC forums

Snigdha Chaturvedi[1], Advisors: Hal Daumé III[1], and Dan Goldwasser[2]

[1]Department of Computer Science, University of Maryland, College Park
[2]Department of Computer Science, Purdue University

Abstract

Finding solutions for the limited personalized interaction between students and instructors in MOOCs (Massive Open Online Course) is fundamental for the success of this emerging trend. Such interactions usually take place on course discussion forums, but the massive scale of these courses limits the instructors' ability to appropriately intervene in relevant discussions. We introduce the problem of identifying relevant intervention points by an instructor in discussion threads. We present predictive models to capture unique aspects of MOOCs, combining course information, forum structure and posts content. Our models can abstract contents of individual posts of threads using latent categories, learned jointly with the binary intervention predictors. Our experiments with data from two *Coursera* MOOCs demonstrate that incorporating the structure of threads leads to highly improved predictive performance.

1 Problem And Motivation

Ubiquitous computing and easy access to high bandwidth internet have reshaped the *modus operandi* in distance education towards Massive Open Online Courses (MOOCs) which impart inexpensive and high-quality education from field-experts to thousands of learners across geographic and cultural barriers.

Even as the MOOC model shows exciting possibilities, it presents a multitude of challenges that must first be negotiated to completely realize its potential. MOOCs platforms have been especially criticized on grounds of lacking a personalized educational experience [12]. Unlike traditional classrooms, the predominant mode of interaction between students and instructors in MOOCs is via online discussion forums which enable students to ask questions and clarify doubts. However, owing to extremely skewed ratios of students to instructional staff, it can be prohibitively time-consuming for the instructional staff to manually follow all threads of a forum. Hence there is a pressing need for automatically *curating* the discussions for the instructors.

Analyzing forum-postings contents and bringing the most pertinent content to the instructor's attention would help instructors receive timely feedback and design interventions as needed. Examining existing forum content demonstrates the severity of the problem from the students' perspective as well. Fig. 1 provides some examples of the ad hoc solution commonly used by MOOC students trying to get instructor's attention —by requesting others to 'up-vote' their posts. existing forum content, indicating that if students want instructor's input on some issues, the only way for them to get his/her attention is by 'up-voting' their votes. Fig. 1 provides some examples of this behavior. This is clearly an

In this paper, we focus on identifying situations in which instructor (used interchangeably with "instructional staff" in this paper) intervention is warranted. We frame identifying instructors intervention in forum threads as a binary prediction problem.

Our initial attempts involved a flat logistic regression classifier based on high level information about the threads and posts. However, forum threads have an inherent structure —the posts of the threads are arranged chronologically in form of a 'chain of events' such that each post is posted in reply to previous ones. The flat classifier couldn't incorporate this structure. Our main technical contribution addresses this issue. We propose two structured models that utilize this behavior. Our models assume that individual posts belong to latent categories which abstractly represent their textual content and that an instructor's decision to reply on a thread is based on this chain of events (represented by the latent categories). Our models present two different ways of modeling instructor's reply decision which can be either simply modeled as the 'next step' in this chain of events (*Linear Chain Markov Model*) or as a decision globally depending on the entire chain (*Global Chain Model*). Our experiments on two different datasets reveal that using the latent post categories helps in better prediction.

In the rest of this paper we describe our approach, contributions and results. For a more detailed description and analysis can be found in [9].

2 Background and Related Work

To the best of our knowledge, the problem of predicting instructor's intervention in MOOC forums has not been

“Problem summary: Anyone else having problems viewing the video lecture...very choppy. If you are also experiencing this issue; please upvote this post.”
“I read that by up-voting threads and posts you can get the instructors’ attention faster.”
“Its is very bad to me that I achieved 10 marks in my 1st assignment and now 9 marks in my 2nd assignment, now I won’t get certificate, please Course staff it is my appeal to change the passing scheme or please be lenient. Please upvote my post so that staff take this problem under consideration.”

Figure 1: Sample posts showing students desiring instructor’s attention have to resolve to the inefficient method of getting their posts upvoted.

addressed yet. Prior work analyzes general online discussion forums of social media sites [18, 4, 11, 2, 19, 5].

Discussion forums have also been used to address other interesting challenges such as extracting chatbox knowledge for use in general online forums [16], automatically extracting answers from discussion forums [7], and subjectivity analysis of forums [6]. Most of these methods use ideas similar to ours: identifying that threads (or discussions) have an underlying structure and that messages belong to categories. However, they operate in a different domain, which makes their goals and methods different from ours.

Our work is most closely related to that of [4] which predicts whether a user who has participated in a thread will later contribute another comment to it. Their prediction task, focusing on users who have already commented on a thread, and their algorithmic approach are different than ours. Our work is also very closely related to that of [27] who predict solvedness —which predicts if there is a solution to the original problem posted in the thread. Like us, they believe that category of posts can assist in the prediction task, however, possibly owing to the complexity of general discussion forums, they had to manually create and annotate data with a sophisticated taxonomy. We do not make such assumptions.

The work presented in [14, 20, 26, 3] discuss characterizing threads using reply-graphs (often trees) and learning this structure. However, this representation is not natural for the MOOC domain where discussions are relatively more focused on the thread topic and are better organized using sections within the forums.

[25] propose a framework for categorizing MOOC forum posts by designing a taxonomy and annotating posts manually to assist general forum analysis. Our model learns categories in a data-driven manner guided by the binary supervision (intervention decision) and serves a different purpose. Nevertheless, in Sec. 4.3 we compare the categories learnt by our models with those proposed by [25].

Apart from this, recent works have looked into interesting challenges in this domain such as better peer grading models [22], code review [15, 21], improving student engagement [1] and understanding how students learn and code [23, 17, 24].

3 Intervention Prediction Models

3.1 Problem Setting

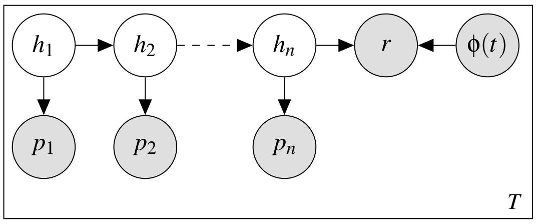
A forum consists of multiple threads. Each thread (t) has a title and consists of multiple posts (p_i). Individual posts do not have a title and the number of posts varies dramatically from one thread to another. We address the problem of predicting if the course instructor would intervene on a thread, t. The instructor’s decision to intervene, r, equals 0 when the instructor does not reply to the thread and 1 otherwise. The only supervision given to all the models during training is in form of intervention decision.

3.2 Initial Attempt: a flat classifier (LR)

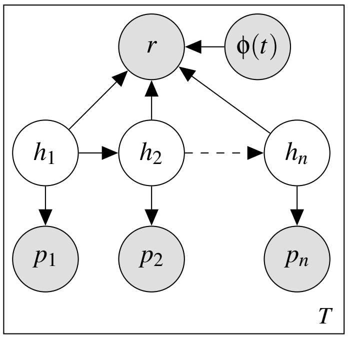
A natural solution to the prediction task calls for a binary classifier. Therefore, we trained a logistic regression classifier which models $P(r|t)$ and uses the following two types of features (collectively referred to as non-structural features). Please note that this is a flat model that does not incorporate thread structure.

Thread only features: capturing high-level information about the thread such as time of first and last postings; number of posts in the thread; binary features indicating whether it was started by anonymous user or if it was marked as approved, unresolved or deleted; and lexical features indicating presence of words such as *assignment*, *exam*, *grade*, *project*, *lecture* in the tile.

Aggregated post features: which analyze the contents of the posts of the thread (compressed into one to maintain a manageable feature space). These included number of votes; statistics about posting times and also their difference from closest assignment/project deadline; lexical features counting occurrences of assessment related words e.g. *grade*, *exam*, thread conclusive words like *thank you*, words indicating technical problems e.g. *submit*, *error* etc.



(a) Linear Chain Markov Model (LCMM)



(b) Global Chain Model (GCM)

Figure 2: Diagrams of the Linear Chain Markov Model (LCMM) and the Global Chain Model (GCM). p_i, r and $\phi(t)$ are observed and h_i are the latent variables. p_i and h_i represent the posts of the thread and their latent categories respectively; r represents the instructor's intervention and $\phi(t)$ represent the non-structural features used by the logistic regression model.

Post Emission Features:

1. $\phi(p_i, h_i)$ = count of occurrences of question words in p_i if the state is h_i; 0 otherwise.
2. $\phi(p_i, h_i)$ = count of occurrences of thank words in p_i if the state is h_i; 0 otherwise.
3. $\phi(p_i, h_i)$ = count of occurrences of greeting words in p_i if the state is h_i; 0 otherwise.
4. $\phi(p_i, h_i)$ = count of occurrences of assessment related words (*grade*, *exam*) in p_i if the state is h_i; 0 otherwise.
5. $\phi(p_i, h_i)$ = count of occurrences of *request*, *submit* or *suggest* in p_i if the state is h_i; 0 otherwise.
6. $\phi(p_i, h_i)$ = log(course duration/$t(p_i)$) if the state is h_i; 0 otherwise. Here $t(p_i)$ is the difference between the posting time of p_i and the closest project/assignment deadline.
7. $\phi(p_i, p_{i-1}, h_i)$ = difference between posting times of p_i and p_{i-1} if the state is h_i; 0 otherwise.

Transition Features:

1. $\phi(h_{i-1}, h_i) = 1$ if previous state is h_{i-1} and current state is h_i; 0 otherwise.
2. $\phi(h_{i-1}, h_i, p_i, p_{i-1})$ = cosine similarity between p_{i-1} and p_i if previous state is h_{i-1} and current state is h_i; 0 otherwise.
3. $\phi(h_{i-1}, h_i, p_i, p_{i-1})$ = length of p_i if previous state is h_{i-1}, p_{i-1} has non-zero question words and current state is h_i; 0 otherwise.
4. $\phi(h_n, r) = 1$ if last post's state is h_n and intervention decision is r; 0 otherwise.
5. $\phi(h_n, r, p_n) = 1$ if last post's state is h_n, p_n has non-zero question words and intervention decision is r; 0 otherwise.
6. $\phi(h_n, r, p_n)$ = log(course duration/$t(p_n)$) if last post's state is h_n and intervention decision is r; 0 otherwise. Here $t(p_n)$ is the difference between the posting time of p_n and the closest project/assignment deadline.

Figure 3: Features used by LCMM and GCM models

3.3 Linear Chain Markov Model (LCMM)

As noted before, posts of a thread are arranged chronologically such that a post is published in reply to the preceding posts and this might effect an instructor's decision to reply. We, therefore, propose to model this complete process using a linear chain markov model shown in Fig. 2a. The model abstractly represents the information from individual posts (p_i) using latent categories (h_i). The intervention decision, r, is the last step in the chain and thus incorporates information from the individual posts.

Apart from non-structural features, this model relies on hand-crafted *emission* and *transition* features (described in Fig. 3) that get fired when the model assigns a latent variable to a post or moves from one latent state to another respectively. The fired features are then multiplied by respective weights to compute a thread's 'score':

$$f_w(t, p) = \max_h [\mathbf{w} \cdot \boldsymbol{\phi(p, r, h, t)}] \quad (1)$$

During training we maximize the combined scores of all threads using a generic EM style iterative two-step algorithm. In the first step, the model uses viterbi algorithm to decode thread sequences with the current weights w_t to find optimal highest scoring latent state sequences that agree with the observed intervention state ($r = r'$). In the next step, given the latent state assignments from the previous step, a structured perceptron algorithm [10] is used to update the weights w_{t+1}.

During prediction, we use the learned weights and viterbi decoding to compute the intervention state and the best scoring latent category sequence.

3.4 Global Chain Model (GCM)

This model (shown in Fig. 2b) proposes another way of incorporating the chain structure of threads while assuming that posts belong to latent categories. It assumes that instructor's decision to intervene is dependent on the latent category of not just the last post but all the posts in the threads. It also uses the features described in Fig. 3.

Similar to the traditional maximum margin based Support Vector Machine (SVM) formulation, our model's objective function is defined as:

$$\min_w \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_j^T l(-r_j f_w(t_j, p_j)) \quad (2)$$

where λ is the regularization coefficient, t_j is the j^{th} thread with intervention decision r_j and p_j are the posts of this thread. $\mathbf{w}$ is the weight vector, $l(\cdot)$ is the squared hinge loss function and $f_w(t_j, p_j)$ is defined in Equation 1.

Replacing the term $f_w(t_j, p_j)$ with it definition in the minimization objective above, reveals the key difference from the traditional SVM formulation - the objective function has a maximum term inside the global minimization problem making it non-convex. We, therefore, employ the optimization algorithm presented in [8] to solve this problem. Exploiting the semi-convexity property [13], the iterative algorithm works in two steps. In the first step, it determines the latent variable assignments for positive examples. In the second step it performs two sub-step iteratively - first it determines the structural assignments for the negative examples, and then optimizes the fixed objective function using a cutting plane algorithm. This is done till it converges for negative examples after which it goes back to the first step. A somewhat similar approach is presented by [28].

At prediction time, given a thread, t, and it posts, p, we use the learned weights to compute $f_w(t, p)$ and classify it as belonging to the positive class (instructor intervenes) if $f_w(t, p) \geq 0$.

4 Empirical Evaluation

4.1 Datasets

For our experiments, we have used the forum content of two MOOCs from different domains (science and humanities): Genes and the Human Condition (GHC) [1] and Women and the Civil Rights Movement (WCR) [2]. Both courses (described in adjacent table) were offered by *Coursera*[3] and taught by professors from the University of Maryland, College Park.

No. of	GHC	WCR
students	30000	14600
instructors	2	1
TAs	3	6
threads	980	800
posts	3,800	3,900

It was commonly observed that after an instructor intervenes on a thread, its posting and/or viewing behavior increases. We, therefore, only consider the student posts until the instructor's first intervention. Also, care was also taken to not use features that changed disproportionately because of the instructor's intervention such as number of views or votes of a thread.

[1] https://www.coursera.org/course/genes

[2] https://www.coursera.org/course/womencivilrights

[3] https://www.coursera.org/

Model	GHC			WCR		
	P	R	F	P	R	F
LR	44.44	16.67	24.24	66.67	15.38	25.00
J48	45.50	20.80	28.55	25.00	23.10	24.01
LCMM	33.33	29.17	31.11	42.86	23.08	**30.00**
GCM	60.00	25.00	**35.29**	50.00	18.52	27.03

Table 1: Held-out test set performances of chain models, LCMM and GCM, are better than that of the flat models, LR and J48.

In our evaluation we approximate instructor's 'should reply' instances with those where the instructor indeed replied. Unlike general forum users, we believe that the correlation between the two scenarios is quite high for instructors. It is their responsibility to reply, and the relatively smaller class sizes of these MOOCs also ensured that most threads were manually reviewed, thus reducing instances of 'missed' threads while retaining the posting behavior and content of a typical MOOC.

Since the purpose of solving this problem is to identify the threads which should be brought to the notice of the instructors, we measure the model-performances using F-measure of the positive class on a held-out test set(randomly selected 20% threads).

4.2 Experimental Results

Table 1 compares the performances of various models on the test set. Parameters values were selected using 10-fold Cross Validation on the training set.

We can see that the chain based models, Linear Chain Markov Model (LCMM) and Global Chain Model (GCM), outperform the flat models (Logistic regression (LR) and Decision Trees (J48)). This validates our hypothesis that using the post structure results in better modeling of instructor's intervention. The table also reveals that GCM yields high precision and low recall values, which is possibly due to the model being more conservative owing to information from all posts of the thread.

4.3 Visual Exploration of Categories

Since our models discover latent categories in a data driven manner, it would be interesting to examine the contents of these categories. Fig. 4 presents a heat map of lexical content of categories identified by LCMM from the GHC dataset. The model discovered 4 categories, which correspond to the rows of the heat map. Its columns represent values of individual features, $f(w, c)$, defined as: $f(w, c) = \frac{C(w,c)}{<C(w,c)>}$ where, $C(w, c)$ is total count of occurrences of a word, w, in

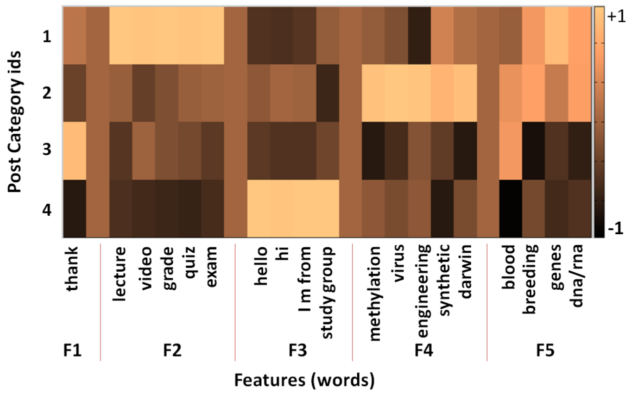


Figure 4: Visualization of lexical contents of the categories learnt by our model from the GHC dataset. Each row is a category and each column represents a feature vector. Brighter colors represent higher values. Categories 1,2,3 and 4 are dominated by Feature group F2, F4, F1 and F3 respectively indicating a semantic segregation of posts by our model's categories.

Category	Example posts
1	'I'm having some issues with video playback. I have downloaded the videos to my laptop...'
2	'In the lecture, she said there are ...I don't see how tumor-suppressor genes are a cancer group mutation.'
3	'Great glossary! Thank you!'
4	'Hi, my name is ... this is my third class with coursera'

Table 2: Representative posts from the four categories learnt by our model.

all posts assigned to category, c and $< C(w, c) >$ represents its expected count based on its frequency in the dataset. We use only a small subset of words in our feature vector for this visualization. These feature values, after normalization, are represented in the heat map using colors ranging from bright cream (high value) to dark black (low value). Thus a darker cell represents lower value.

For visual convenience, the features are manually clustered into five groups (F1 to F5) each separated by a dark beige colored column. Group F1consists of words like *thank you, thanks* etc. which are characteristic of posts that mark the conclusion of a resolved thread. Row corresponding to the category 3 in Table 2 shows example of such a post. Similarly, F2 represents the features related to logistics of the course and F3 captures introductory posts by new students. Finally, F4 contains words would appear in posts that discuss specific aspects of the course contents, while F5 contains general buzz words that would appear frequently in any biology course.

Analyzing the heat map, we can see that Categories 1, 2, 3 and 4 are dominated by logistics (F2), course content related (F4), thank you (F1) and introductory posts (F3) respectively, represented by bright colors in their respective rows. We also observe similar correlations while examining the columns of the heat map. Also, F5, which contains words common to the gene and human health domain, is scattered across multiple categories.

Table 2 gives examples of representative posts from the four clusters. Due to space constraints, we show only part of the post. We can see that these examples agree with our observations from the heat map.

As noted in Sec. 2, we compare the semantics of clusters learnt by our models with those proposed by [25] even though the two categorizations are not directly comparable. Nevertheless, generally speaking, our category 1 corresponds to [25]'s *Course structure/policies* and category 2 corresponds to *Content*. Interestingly, categories 3 and 4, correspond to a single *Social/affective* from the previous work.

Thus, the model splits posts into semantically coherent categories, despite the lack of any explicit training data, which correspond well with taxonomies for classifying forums posts defined by domain experts [25].

5 Contributions

The massive scale of MOOCs limits personalized interaction, and needs instructors to browse forum threads, and selectively respond when appropriate. This time consuming and error prone task stresses the need for automatically supplying this actionable information. This paper takes first steps in this direction, and formulates the novel problem of predicting instructor intervention in MOOC forums.

Our experiments on forum data from two different *Coursera* MOOCs show that utilizing thread structure is important for predicting instructor's behavior. Furthermore, our qualitative analysis shows that our latent categories are semantically coherent to human eye.

Our contributions can be summarized as:

- We motivate and introduce the problem of predicting instructor intervention in MOOC forums
- We present two chain-based models that incorporate thread structure by abstractly representing posts using latent variables.
- We show the utility of modeling thread structure, and the value of lexical and domain specific knowledge for the prediction task

References

[1] A. Anderson, D. P. Huttenlocher, J. M. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *WWW*, pages 687–698, 2014.

[2] Y. Artzi, P. Pantel, and M. Gamon. Predicting responses to microblog posts. In *NAACL HLT*, pages 602–606, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] E. Aumayr, J. Chan, and C. Hayes. Reconstruction of threaded conversations in online discussion forums. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.

[4] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *WSDM*, pages 13–22, New York, NY, USA, 2013. ACM.

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *WSDM*, pages 65–74, New York, NY, USA, 2011. ACM.

[6] P. Biyani, C. Caragea, and P. Mitra. Predicting subjectivity orientation of online forum threads. In *CICLing (2)*, pages 109–120, 2013.

[7] R. Catherine, R. Gangadharaiah, K. Visweswariah, and D. Raghu. Semi-supervised answer extraction from discussion forums. In *IJCNLP*, pages 1–9, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[8] M.-W. Chang, D. Goldwasser, D. Roth, and V. Srikumar. Discriminative learning over constrained latent representations. In *NAACL HLT*, pages 429–437, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[9] S. Chaturvedi, D. Goldwasser, and H. Daumé III. Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1501–1511. Association for Computational Linguistics, 2014.

[10] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *EMNLP*, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[11] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. What makes conversations interesting? themes, participants and consequences of conversations in online social media. In *WWW*, pages 331–331, April 2009.

[12] M. Edmundson. The trouble with online education, July 19 2012. `http://www.nytimes.com/2012/07/20/opinion/the-trouble-with-online-education.html`.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[14] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW*.

[15] J. Huang, C. Piech, A. Nguyen, and L. J. Guibas. Syntactic and functional variability of a million code submissions in a machine learning mooc. In *AIED Workshops*, 2013.

[16] J. Huang, M. Zhou, and D. Yang. Extracting chatbox knowledge from online discussion forums. In *IJCAI*, pages 423–428, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

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

[18] J. M. Kleinberg. Computational perspectives on social phenomena at global scales. In F. Rossi, editor, *IJCAI*. IJCAI/AAAI, 2013.

[19] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, New York, NY, USA, 2010. ACM.

[20] D. Liben-Nowell and J. Kleinberg. Tracing the flow of information on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 25 Mar. 2008.

[21] A. Nguyen, C. Piech, J. Huang, and L. J. Guibas. Codewebs: scalable homework search for massive open online programming courses. In *WWW*, pages 491–502, 2014.

[22] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. In *Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013)*, 2013.

[23] C. Piech, M. Sahami, D. Koller, S. Cooper, and P. Blikstein. Modeling how students learn to program. In *SIGCSE*, pages 153–160, 2012.

[24] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, and L. Getoor. Modeling learner engagement in moocs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*, 2013.

[25] G. S. Stump, J. DeBoer, J. Whittinghill, and L. Breslow. Development of a framework to classify mooc discussion forum posts: Methodology and challenges. 2013.

[26] H. Wang, C. Wang, C. Zhai, and J. Han. Learning online discussion structures by conditional random fields. In *SIGIR*, pages 435–444, New York, NY, USA, 2011. ACM.

[27] L. Wang, S. N. Kim, and T. Baldwin. The utility of discourse structure in forum thread retrieval. In *AIRS*, pages 284–295, 2013.

[28] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176, New York, NY, USA, 2009. ACM.