Predicting Instructor's Intervention in MOOC forums

Snigdha Chaturvedi **Dan Goldwasser** **Hal Daumé III**
Department of Computer Science,
University of Maryland, College Park, Maryland
{snigdhac, goldwas1, hal}@umiacs.umd.edu

Abstract

Instructor intervention in student discussion forums is a vital component in Massive Open Online Courses (MOOCs), where personalized interaction is limited. This paper introduces the problem of predicting instructor interventions in MOOC forums. We propose several prediction models designed to capture unique aspects of MOOCs, combining course information, forum structure and posts content. Our models abstract contents of individual posts of threads using latent categories, learned jointly with the binary intervention prediction problem. Experiments over data from two *Coursera* MOOCs demonstrate that incorporating the structure of threads into the learning problem leads to better predictive performance.

1 Introduction

Ubiquitous computing and easy access to high bandwidth internet have reshaped the *modus operandi* in distance education towards Massive Open Online Courses (MOOCs). Courses offered by ventures such as Coursera and Udacity now 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 (Edmundson, 2012). Unlike traditional classrooms, the predominant mode of interaction between students and instructors in MOOCs is via online discussion forums. Ideally, forum discussions can help make up for the lack of direct interaction, by enabling students to ask questions and clarify doubts. However, due to huge class sizes, even during the short duration of a course, MOOCs witness a very large number of threads on these forums. 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.

In this paper, we focus on identifying situations in which instructor (used interchangeably with "instructional staff" in this paper) intervention is warranted. Using existing forum posts and interactions, we frame this as a binary prediction problem of identifying instructor's intervention in forum threads. Our initial analysis revealed that instructors usually intervene on threads discussing students' issues close to a quiz or exam. They also take interest in grading issues and logistics problems. There are multiple cues specific to the MOOC setting, which when combined with the rich lexical information present in the forums, can yield useful predictive models.

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. From the students' perspective, the problem is evident from an examination of 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 inefficient solution.

Our main technical contribution is introducing three different models addressing the task of predicting instructor interventions. The first uses a logistic regression model that primarily incorporates high level information about threads and posts. However, forum threads have structure which is not leveraged our initial model. We present two

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1501–1511,
Baltimore, Maryland, USA, June 23-25 2014.

"The 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 that showing students desiring instructor's attention have to resolve to the inefficient method of getting their posts upvoted.

additional structured models. Both models assume that posts of a thread structure it in form of a story or a "chain of events." For example, an opening post of a thread might pose a question and the following posts can then answer or comment on the question. Our second and third models tap this linear 'chain of events' behavior by assuming that individual posts belong to latent categories which represent their textual content at an abstract level and that an instructor's decision to reply to a post is based on this chain of events (represented by the latent categories). We present two different ways of utilizing this 'chain of events' behavior for predicting instructor's intervention which can be either simply modeled as the 'next step' is 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.

Our contributions can be summarized as:

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

2 Related Work

To the best of our knowledge, the problem of predicting instructor's intervention in MOOC forums has not been addressed yet. Prior work deals with analyzing general online discussion forums of social media sites (Kleinberg, 2013): such as predicting comment volume (Backstrom et al., 2013; De Choudhury et al., 2009; Wang et al., 2012; Tsagkias et al., 2009; Yano and Smith, 2010; Artzi et al., 2012) and rate of content diffusion (Kwak et al., 2010; Lerman and Ghosh, 2010; Bakshy et al., 2011; Romero et al., 2011; Artzi et al., 2012) and also question answering (Chaturvedi et al., 2014).

Wang et al. (2007) incorporate thread structure of conversations using features in email threads while Goldwasser and Daumé III (2014) use latent structure, aimed to identify relevant dialog segments, for predicting objections during courtroom deliberations. Other related work include speech act recognition in emails and forums but at a sentence level (Jeong et al., 2009), and using social network analysis to improve message classification into pre-determined types (Fortuna et al., 2007). Discussion forums data has also been used to address other interesting challenges such as extracting chatbox knowledge for use in general online forums (Huang et al., 2007) and automatically extracting answers from discussion forums (Catherine et al., 2013), subjectivity analysis of online forums (Biyani et al., 2013). 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 Backstrom et al. (2013) which introduced the re-entry prediction task —predicting whether a user who has participated in a thread will later contribute another comment to it. While seemingly related, 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 Wang et al. (2013) 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 (Gómez et al., 2008;

Liben-Nowell and Kleinberg, 2008; Kumar et al., 2010; Golub and Jackson, 2010; Wang et al., 2011; Aumayr et al., 2011) 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.

Although most prior work focuses on discussion forums of social media sites such as Twitter or Facebook, where the dynamics of interaction is very different from MOOCs, a small number of recent work address the unique MOOC setting.

Stump et al. (2013) propose a framework for categorizing 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 Stump et al. (2013).

Apart from this, recent works have looked into interesting challenges in this domain such as better peer grading models (Piech et al., 2013), code review (Huang et al., 2013; Nguyen et al., 2014), improving student engagement (Anderson et al., 2014) and understanding how students learn and code (Piech et al., 2012; Kizilcec et al., 2013; Ramesh et al., 2013).

3 Intervention Prediction Models

In this section, we explain our models in detail.

3.1 Problem Setting

In our description it is assumed that a discussion board is organized into multiple forums (representing topics such as "Assignment", "Study Group" etc.). 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 doesn't reply to the thread and 1 otherwise. The individual posts are not assumed to be labeled with any category and the only supervision given to the model during training is in form of intervention decision.

3.2 Logistic Regression (LR)

Our first attempt at solving this problem involved training a logistic regression for the binary prediction task which models $P(r|t)$.

3.2.1 Feature Engineering

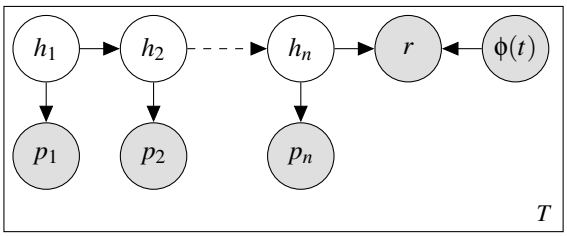
Our logistic regression model uses the following two types of features: *Thread only features* and *Aggregated post features*. 'Thread only features' capture information about the thread such as when, where, by who was the thread posted and lexical features based on the title of the thread. While these features provide a high-level information about the thread, it is also important to analyze the contents of the posts of the thread. In order to maintain a manageable feature space, we compress the features from posts and represent them using our 'Aggregated post features'.

Thread only features:

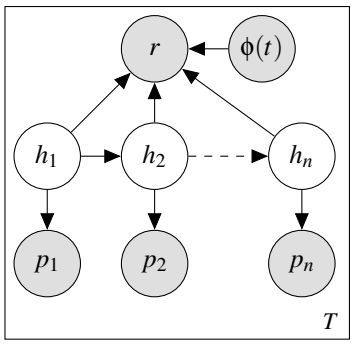
1. a binary feature indicating if the thread was started by an anonymous user
2. three binary features indicating whether the thread was marked as approved, unresolved or deleted (respectively)
3. forum id in which the thread was posted
4. time when the thread was started
5. time of last posting on the thread
6. total number of posts in the thread
7. a binary feature indicating if the thread title contains the words *lecture* or *lectures*
8. a binary feature indicating if the thread title contains the words *assignment*, *quiz*, *grade*, *project*, *exam* (and their plural forms)

Aggregated post features:

9. sum of number of votes received by the individual posts
10. mean and variance of the posting times of individual posts in the thread
11. mean of time difference between the posting times of individual posts and the closest course landmark. A course landmark is the deadline of an assignment, exam or project.
12. sum of count of occurrences of assessment related words e.g. *grade*, *exam*, *assignment*, *quiz*, *reading*, *project* etc. in the posts
13. sum of count of occurrences of words indicating technical problems e.g. *problem*, *error*
14. sum of count of occurrences of thread conclusive words like *thank you* and *thank*
15. sum of count of occurrences of *request*, *submit*, *suggest*



(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.

We had also considered and dropped (because of no performance gain) other features about identity of the user who started the thread, number of distinct participants in the thread (an important feature used by Backstrom et al. (2013)), binary feature indicating if the first and the last posts were by the same user, average number of words in the thread's posts, lexical features capturing references to the instructors in the posts etc.

3.3 Linear Chain Markov Model (LCMM)

The logistic regression model is good at exploiting the thread level features but not the content of individual posts. The 'Aggregated post features' attempt to capture this information but since the number of posts in a thread is variable, these features relied on aggregated values. We believe that considering aggregate values is not sufficient for the task in hand. As noted before, posts of a thread are not independent of each other. Instead, they 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. For example, consider a thread that starts with a question. The following posts will be students' attempt to answer the question or raise further concerns or comment on previous posts. The instructor's post, though a future event, will be a part of this process.

For every thread, t, in the dataset:

1. Choose a start state, h_1, and emit the first post, p_1.
2. For every subsequent post, $p_i \ \forall \ i \ \in \{2 \dots n\}$:
   (a) Transition from h_{i-1} to h_i.
   (b) Emit post p_i.
3. Generate the instructor's intervention decision, r, using the last state h_n and non-structural features, $\phi(t)$.

Figure 3: Instructor's intervention decision process for the Linear Chain Markov Model.

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. It also depends on the thread level features: 'Thread only features' and the 'Aggregated post features' jointly represented by $\phi(t)$ (also referred to as the non-structural features). This process is explained in Fig. 3.

We use hand-crafted features to model the dynamics of the generative process. Whenever a latent state emits a post or transits to another latent state (or to the final intervention decision state), emission and transition features get fired which are then multiplied by respective weights to compute a thread's 'score':

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

Note that the non-structural features, $\phi(t)$, also contribute to the final score.

3.3.1 Learning and Inference

During training we maximize the combined scores of all threads in the dataset using a generic EM style algorithm. The supervision in this model is provided only in form of the observed intervention decision, r and the post categories, h_i are hid-

den. The model uses the pseudocode shown in Algorithm 1 to iteratively refine the weight vectors. In each iteration, the model first 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 (Collins, 2002) is used to update the weights w_{t+1} using weights from the previous step, w_t, initialization.

Algorithm 1 Training algorithm for LCMM

1: **Input:** Labeled data $D = \{(t, p, r)_i\}$
2: **Output:** Weights w
3: **Initialization:** Set w_j randomly, $\forall j$
4: **for** $t : 1$ to N **do**
5: $\hat{h_i} = \arg\max_h[\mathbf{w_t} \cdot \boldsymbol{\phi(p, r, h, t)}]$ such that $r = r_i \forall i$
6: w_{t+1} = StructuredPerceptron$(t, p, \hat{h}, r)$
7: **end for**
8: return w

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

3.3.2 Feature Engineering

In addition to the 'Thread Only Features' and the 'Aggregated post features', $\phi(t)$ (Sec. 3.2.1, this model uses the following emission and transition features:

Post Emission Features:

1. $\phi(p_i, h_i)$ = count of occurrences of question words or question marks in p_i if the state is h_i; 0 otherwise.
2. $\phi(p_i, h_i)$ = count of occurrences of thank words (*thank you* or *thanks*) in p_i if the state is h_i; 0 otherwise.
3. $\phi(p_i, h_i)$ = count of occurrences of greeting words (e.g. *hi*, *hello*, *good morning*, *welcome* etc) in p_i if the state is h_i; 0 otherwise.
4. $\phi(p_i, h_i)$ = count of occurrences of assessment related words (e.g. *grade*, *exam*, *assignment*, *quiz*, *reading*, *project* etc.) 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 course landmark (assignment or project deadline or exam).
7. $\phi(p_i, p_{i-1}, h_i)$ = difference between posting times of p_i and p_{i-1} normalized by course duration 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 course landmark (assignment or project deadline or exam).

3.4 Global Chain Model (GCM)

In this model we propose another way of incorporating the chain structure of a thread. Like the previous model, this model also assumes that posts belong to latent categories. It, however, doesn't model the instructor's intervention decision as a step in the thread generation process. Instead, it assumes that instructor's decision to intervene is dependent on all the posts in the threads, modeled using the latent post categories. This model is shown in Fig. 2b. Assuming that p represents posts of thread t, h represents the latent category assignments, r represents the intervention decision; feature vector, $\boldsymbol{\phi(p, r, h, t)}$, is extracted for each thread and using the weight vector, $\mathbf{w}$, this model defines a decision function, similar to what is shown in Equation 1.

3.4.1 Learning and Inference

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 the contents of Equation 1 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 (Chang et al., 2010) to solve this problem. Exploiting the semi-convexity property (Felzenszwalb et al., 2010), the algorithm works in two steps, each executed iteratively. In the first step, it determines the latent variable assignments for positive examples. The algorithm then performs two step iteratively - first it determines the structural assignments for the negative examples, and then optimizes the fixed objective function using a cutting plane algorithm. Once this process converges for negative examples, the algorithm reassigns values to the latent variables for positive examples, and proceeds to the second step. The algorithm stops once a local minimum is reached. A somewhat similar approach, which uses the Convex-Concave Procedure (CCCP) is presented by (Yu and Joachims, 2009).

At test 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 $f_w(t, p) \geq 0$.

3.4.2 Feature Engineering

The feature set used by this model is very similar to the features used by the previous model. In addition to the non-structural features used by the logistic regression model (Sec. 3.2.1), it uses all the Post Emission features and the three transition features represented by $\phi(h_{i-1}, h_i)$ and $\phi(h_{i-1}, h_i, p_i, p_{i-1})$ as described in Sec. 3.3.2.

4 Empirical Evaluation

This section describes our experiments.

4.1 Datasets and Evaluation Measure

For our experiments, we have used the forum content of two MOOCs from different domains (science and humanities), offered by *Coursera*[1], a leading education technology company. Both courses were taught by professors from the University of Maryland, College Park.

Genes and the Human Condition (From Behavior to Biotechnology) (GHC) dataset. [2] This course was attended by 30,000 students and the instructional staff comprised of 2 instructors, 3 Teaching Assistants and 56 technical support staff. The discussion forum of this course consisted of 980 threads composed of about 3,800 posts.

Women and the Civil Rights Movement (WCR) dataset. [3] The course consisted of a classroom of about 14,600 students, 1 instructor, 6 Teaching Assistants and 49 support staff. Its discussion forum consisted of 800 threads and 3,900 posts.

We evaluate our models on held-out test sets. For the GHC dataset, the test set consisted of 186 threads out of which the instructor intervened on 24 while, for the WCR dataset, the instructor intervened on 21 out of 155 threads.

Also, 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. Care was also taken to not use features that increased/decreased disproportionately because of the instructor's intervention such as number of views or votes of a thread.

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 by choosing to a MOOC, they have 'bought in' to the idea of forum participation. The relatively smaller class sizes of these two 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.

4.2 Experimental Results

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 performance of our models using F-measure of the positive class. The values of various parameters were selected using 10-fold Cross Validation on

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

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

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

Model	Genes and the Human Condition (GHC)			Women and the Civil Rights (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 unstructured models, LR and J48.

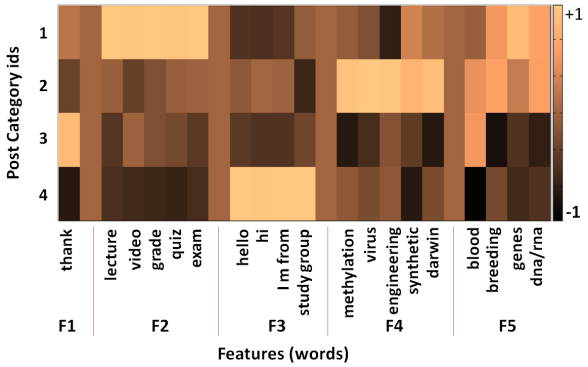


Figure 4: Visualization of lexical contents of the categories learnt by our model from the GHC dataset. Each row is a category and each column represents a feature vector. Bright cream color represents high values while lower values are represented by darker shades. Dark beige columns are used to better separate the five feature clusters, F1-F5, which represent words that are common in thanking, logistics-related, introductory, syllabus related and miscellaneous posts respectively. Categories 1,2,3 and 4 are dominated by F2, F4, F1 and F3 respectively indicating a semantic segregation of posts by our model's categories.

the training set. Table 1 presents the performances of the proposed models on the held-out test sets. We also report performance of a decision tree (J48) on the test sets for sake of comparison.

We can see that the chain based models, Linear Chain Markov Model (LCMM) and Global Chain Model (GCM), outperform the unstructured models, namely 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

Our chain based models assume that posts belong to different (latent) categories and use these categories to make intervention predictions. Since this process of discovering categories is data driven, 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 value of H (number of categories) was set to be 4 and was predetermined during the model selection procedure. Each row of the heat map represents a category and the 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 all posts assigned to category, c and $< C(w, c) >$ represents its expected count based on its frequency in the dataset. While the actual size of vocabulary is huge, 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). The darker the shade of a cell, the lower is the value represented by it.

For visual convenience, the features are manually clustered into five groups (F1 to F5) each separated by a dark beige colored column in the heat map. The first column of the heat map represents the F1 group which consists of words like *thank you, thanks* etc. These words are characteristic of posts that mark either the conclusion of a resolved thread or are posted towards the end of the course. Rows corresponding to the category 3 in Table 2 show two examples of such posts. Similarly, F2 represents the features related to logistics of the course and F3 captures introductory posts by new students. Finally, F4 contains words that are closely related to the subfield of gene and human conditions and would appear in posts that discuss specific aspects or chapters of the course con-

tents, while F5 contains general buzz words that would appear frequently in any biology course.

Analyzing individual rows of the heat map, we can see that out of F1 to F4, 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. For example, *dna/rna* and *breeding* are sufficiently frequent in category 1 as well as 2.

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

Furthermore, as noted in Sec. 2, we compare the semantics of clusters learnt by our models with those proposed by Stump et al. (2013) even though the two categorizations are not directly comparable. Nevertheless, generally speaking, our category 1 corresponds to Stump et al. (2013)'s *Course structure/policies* and category 2 corresponds to *Content*. Interestingly, categories 3 and 4, which represent valedictory and introductory posts, correspond to a single *Social/affective* from the previous work.

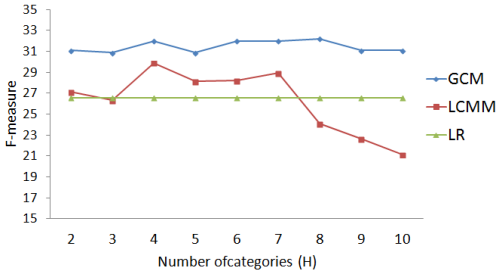
We can, therefore, conclude that the model, indeed splits the posts into categories that look semantically coherent to the human eyes.

4.4 Choice of Number of Categories

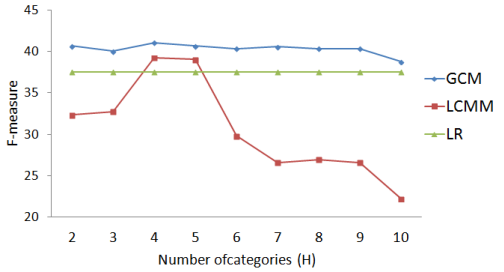
Our chain based models, assigning forum posts to latent categories, are parameterized with H, the number of categories. We therefore, study the sensitivity of our models to this parameter. Fig. 5, plots the 10-fold cross validation performance of the models with increasing values of H for the two datasets. Interestingly, the sensitivity of the two models to the value of H is very different.

The LCMM model's performance fluctuates as the value of H increases. The initial performance improvement might be due to an increase in the expressive power of the model. Performance peaks at $H = 4$ and then decreases, perhaps owing to over-fitting of the data.

In contrast, GCM performance remains steady for various values of H which might be attributed



(a) Genes and the Human Condition dataset



(b) Women and the Civil Rights Movement dataset

Figure 5: Cross validation performances of the two models with increasing number of categories.

to the explicit regularization coefficient which helps combat over-fitting, by encouraging zero weights for unnecessary categories.

4.5 How important are linguistic features?

We now focus on the structure independent features and experiment with their predictive value, according to types. We divide the features used by the LR into the following categories:[4]

- Full: set of all features (feature no. 1 to 15)
- lexical: based on content of thread titles and posts (feature no. 7 to 8 and 12 to 13)
- landmark: based on course landmarks (e.g, exams, quizzes) information (feature no. 11)
- MOOCs-specific: features specific to the MOOCs domain (lexical + landmark features)
- post: based only on aggregated posts information (feature no. 9 to 15)
- temporal: based on posting time patterns (feature no. 4, 5 and 10)

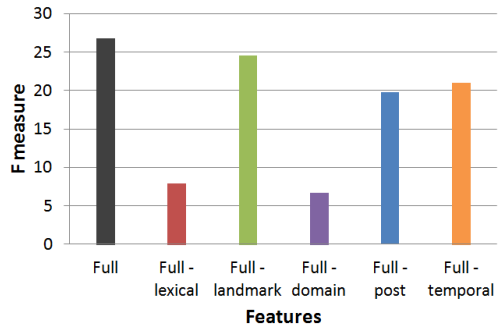
Fig. 6 shows 10-fold cross validation F-measure of the positive class for LR when different types of features are excluded from the full set.

The figure reveals that the MOOCs-specific features (purple bar) are important for both the datasets indicating a need for designing specialized models for forums analysis in this domain.

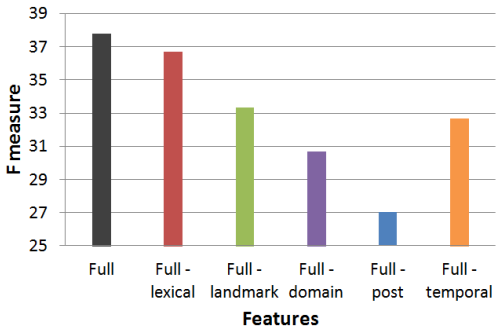
[4]Please refer to Sec 3.2.1 for description of the feature id.

Category	Example posts
1	'I'm having some issues with video playback. I have downloaded the videos to my laptop...'
1	'There was no mention of the nuclear envelope in the Week One lecture, yet it was in the quiz. Is this a mistake?'
2	'DNA methylation is a crucial part of normal development of organisms and cell differentiation in higher organisms...'
2	'In the lecture, she said there are...I don't see how tumor-suppressor genes are a cancer group mutation.'
3	'Thank you very much for a most enjoyable and informative course.'
3	'Great glossary! Thank you!'
4	'Hello everyone, I'm ... from the Netherlands. I'm a life science student.
4	'Hi, my name is ... this is my third class with coursera'

Table 2: Representative posts from the four categories learnt by our model. Due to space and privacy concerns we omit some parts of the text, indicated by ". . . ".



(a) Genes and the Human Condition dataset



(b) Women and the Civil Rights Movement dataset

Figure 6: Cross validation performances of the various feature types for the two datasets.

Also, lexical features (red bar) and post features (blue bar) have pretty dramatic effects in GHC and WCR data respectively.

Interestingly, removing the landmark feature set (green bar) causes a considerable drop in predictive performance, even though it consists of only one feature. Other temporal features (orange bar) also turn out to be important for the prediction. From a separate instructor activity vs time graph (not shown due to space constraints), we observed that instructors tend to get more active as the course progresses and their activity level also increases around quizzes/exams deadlines.

We can, therefore, conclude that all feature types are important and that lexical as well as MOOC specific analysis is necessary for modeling instructor's intervention.

5 Conclusion

One of the main challenges in MOOCs is managing student-instructor interaction. The massive scale of these courses rules out any form of personalized interaction, leaving instructors with the need to go over the forum discussions, gauge student reactions and selectively respond when appropriate. This time consuming and error prone task stresses the need for methods and tools supplying this actionable information automatically.

This paper takes a first step in that direction, and formulates the novel problem of predicting instructor intervention in MOOC discussion forums. Our main technical contribution is to construct predictive models combining information about forum post content and posting behavior with information about the course and its landmarks.

We propose three models for addressing the task. The first, a logistic regression model is trained on thread level and aggregated post features. The other two models take thread structure into account when making the prediction. These models assume that posts can be represented by categories which characterize post content at an abstract level, and treat category assignments as latent variables organized according to, and influenced by, the forum thread structure.

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.

References

Ashton Anderson, Daniel P. Huttenlocher, Jon M. Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *WWW*, pages 687–698.

Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 602–606, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik Aumayr, Jeffrey Chan, and Conor Hayes. 2011. Reconstruction of threaded conversations in online discussion forums. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2013. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 13–22, New York, NY, USA. ACM.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA. ACM.

Prakhar Biyani, Cornelia Caragea, and Prasenjit Mitra. 2013. Predicting subjectivity orientation of online forum threads. In *CICLing (2)*, pages 109–120.

Rose Catherine, Rashmi Gangadharaiah, Karthik Visweswariah, and Dinesh Raghu. 2013. Semi-supervised answer extraction from discussion forums. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1–9, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Ming-Wei Chang, Dan Goldwasser, Dan Roth, and Vivek Srikumar. 2010. Discriminative learning over constrained latent representations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 429–437, Stroudsburg, PA, USA. Association for Computational Linguistics.

Snigdha Chaturvedi, Vittorio Castelli, Radu Florian, Ramesh M. Nallapati, and Hema Raghavan. 2014. Joint question clustering and relevance prediction for open domain non-factoid question answering. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 503–514, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. 2009. What makes conversations interesting? themes, participants and consequences of conversations in online social media. In *18th International World Wide Web Conference (WWW)*, pages 331–331, April.

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

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Blaz Fortuna, Eduarda Mendes Rodrigues, and Natasa Milic-Frayling. 2007. Improving the classification of newsgroup messages through social network analysis. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 877–880, New York, NY, USA. ACM.

Dan Goldwasser and Hal Daumé III. 2014. "I object!" modeling latent pragmatic effects in courtroom dialogues. *European Chapter of the Association for Computational Linguistics (EACL)*, April. To appear.

Benjamin Golub and Matthew O. Jackson. 2010. Seeing only the successes: The power of selection bias in explaining the structure of observed internet diffusions.

Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 645–654, New York, NY, USA. ACM.

Jizhou Huang, Ming Zhou, and Dan Yang. 2007. Extracting chatbox knowledge from online discussion forums. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 423–428, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jonathan Huang, Chris Piech, Andy Nguyen, and Leonidas J. Guibas. 2013. Syntactic and functional variability of a million code submissions in a machine learning mooc. In *AIED Workshops*.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1250–1259, Stroudsburg, PA, USA. Association for Computational Linguistics.

René F. Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *LAK*, pages 170–179.

Jon M. Kleinberg. 2013. Computational perspectives on social phenomena at global scales. In Francesca Rossi, editor, *IJCAI*. IJCAI/AAAI.

Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. 2010. Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 553–562, New York, NY, USA. ACM.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

K. Lerman and R. Ghosh. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*.

David Liben-Nowell and Jon Kleinberg. 2008. 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 March.

Andy Nguyen, Christopher Piech, Jonathan Huang, and Leonidas J. Guibas. 2014. Codewebs: scalable homework search for massive open online programming courses. In *WWW*, pages 491–502.

Chris Piech, Mehran Sahami, Daphne Koller, Steve Cooper, and Paulo Blikstein. 2012. Modeling how students learn to program. In *SIGCSE*, pages 153–160.

Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned models of peer assessment in MOOCs. In *Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013)*.

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

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, USA. ACM.

Glenda S. Stump, Jennifer DeBoer, Jonathan Whittinghill, and Lori Breslow. 2013. Development of a framework to classify mooc discussion forum posts: Methodology and challenges.

Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. 2009. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1765–1768, New York, NY, USA. ACM.

Yi-Chia Wang, Mahesh Joshi, and Carolyn Penstein Ros. 2007. A feature based approach to leveraging context for classifying newsgroup style discussion segments. In John A. Carroll, Antal van den Bosch, and Annie Zaenen, editors, *ACL*. The Association for Computational Linguistics.

Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011. Learning online discussion structures by conditional random fields. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 435–444, New York, NY, USA. ACM.

Chunyan Wang, Mao Ye, and Bernardo A. Huberman. 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 244–252, New York, NY, USA. ACM.

Li Wang, Su Nam Kim, and Timothy Baldwin. 2013. The utility of discourse structure in forum thread retrieval. In *AIRS*, pages 284–295.

Tae Yano and Noah A. Smith. 2010. What's worthy of comment? content and comment volume in political blogs. In William W. Cohen and Samuel Gosling, editors, *ICWSM*. The AAAI Press.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, New York, NY, USA. ACM.