Bringing Order to Chaos in MOOC Discussion Forums with Content-Related Thread Identification

Alyssa Friend Wise
Simon Fraser University
250-13450 102nd Avenue
Surrey, B.C. V3T 0A3 Canada
1-778-782-8046
alyssa_wise@sfu.ca

Yi Cui
Simon Fraser University
250-13450 102nd Avenue
Surrey, B.C. V3T 0A3 Canada
1-778-782-8046
yca231@sfu.ca

Jovita Vytasek
Simon Fraser University
250-13450 102nd Avenue
Surrey, B.C. V3T 0A3 Canada
1-778-782-8046
jvytasek@sfu.ca

ABSTRACT
This study addresses the issues of overload and chaos in MOOC discussion forums by developing a model to categorize and identify threads based on whether or not they are substantially related to the course content. Content-related posts were defined as those that give/seek help for the learning of course material and share/comment on relevant resources. A linguistic model was built based on manually-coded starting posts in threads from a statistics MOOC (n=837) and tested on thread starting posts from the second offering of the same course (n=304) and a different statistics course (n=298). The number of views and votes threads received were tested to see if they helped classification. Results showed that content-related posts in the statistics MOOC had distinct linguistic features which appeared to be unrelated to the subject-matter domain; the linguistic model demonstrated good cross-course reliability (all recall and precision > .77) and was useful across all time segments of the courses; number of views and votes were not helpful for classification.

CCS Concepts
• **Information systems ~ Content analysis and feature selection** • **Information systems ~ Clustering and classification** • **Applied computing ~ E-learning** • Computing methodologies ~ Natural language processing • Computing methodologies ~ Model development and analysis

Keywords
Massive open online courses; social interaction; discussion forum; natural language processing; machine learning

1. INTRODUCTION
Massive Open Online Courses (MOOCs) are online learning environments that are open to anyone with web access [7]. Generally charging no cost for participation and setting no prerequisite requirements, the courses often attract thousands, or even hundreds of thousands of registrations. In the past several years, MOOCs have seen dramatic growth in terms of available platforms, participating institutions, courses offered, and learners involved [10, 28].

With such large scale enrollments and the support of the Internet, MOOCs have the potential to provide abundant interaction opportunities for learners. Interaction is a key component in the quality of online learning, providing important support to learners [31]. Specifically, providing a good interaction environment is considered an important criterion for the quality of MOOCs [17]. While many forms of interaction are theoretically possible, currently discussion forums are the de facto primary venue in MOOCs for leaner-learner and learner-instructor interactions. Discussion forums are valued by instructors as an important instrument for understanding and intervening in learning activities [29, 16], while learners use them for giving and getting help for challenges they encounter in their learning [2, 30].

In order for these activities to happen effectively, instructors and learners need to be able to find the messages relevant to their purposes. However, due to the large number of participants in MOOCs, discussion forums are often plagued by information overload and chaos [24, 3]. On top of this, a large proportion of MOOC posts are not directly related to the course [3]. As a result, forums become overwhelming and confusing for users to navigate [15]. This is an exacerbation of problems seen in traditional discussion forums for over a decade [13, 9, 26]. Such problems can lead to extremely low levels of responsiveness [14] which means the desired understanding, intervening, help giving and getting activities are not optimized [12].

Currently, there are very limited means for addressing overload and disorder in MOOC discussion forums. One commonly used strategy is to set sub-forums for different purposes. However, misplaced posts are common in MOOCs [27], indicating that this strategy hasn't been very successful. An alternative strategy has been to ask learners to tag their posts, making it easier for others to identify different types of postings. But just like with sub-forums, there is no guarantee that learners will tag their posts in accurate and consistent ways. In addition, many MOOC forums allow learners to sort posts by the number of views and votes made by other users. However, such forms of peer recommendation are subject to bias through the disproportionate effect of early support (the "rich get richer" phenomenon) and positioning effects [20]. More importantly, since these features simply indicate the general "popularity" of messages, their value in distinguishing the type of information different posts contain is questionable [6]. This indicates the need for novel tools that can

LAK '16, April 25 - 29, 2016, Edinburgh, United Kingdom
 ACM 978-1-4503-4190-5/16/04…$15.00
DOI: dx.doi.org/10.1145/2883851.2883916

more effectively assist instructors and learners in navigating the complicated landscape of MOOC discussion forums to find the posts they are looking for.

In this study, we address the overload problem in MOOC discussion forums by developing a model to automatically classify threads as substantially related to the course material or not. In the following section we first review prior efforts to address overload in MOOC discussion forums and justify why helping instructors and learners easily distinguish content-related and non-content-related posts is a useful and novel contribution to this problem space. We then describe how the concept was operationalized to manually code threads from three statistics MOOC discussion forums according to whether or not they were substantially related to the course subject, and subsequently develop and test a classification model to automatically identify content-related threads. Through this work, we aim to build a foundation for tools that can help MOOC instructors and learners locate forum threads addressing the learning of course material more easily.

2. LITERATURE REVIEW

Research efforts targeting overload and chaos in MOOC discussion forums generally fall into two categories. One is machine-oriented solutions that use automatic tools to diagnose learner's posts and provide prescribed help or resource as appropriate. For example, [1] targeted the lack of responsiveness in MOOC forums with an automatic tool that detects confusion in posts and recommends relevant video clips to the learner. Automated approaches have value in providing personalized instant intervention to learners in need of simple answers to straightforward questions; however, in many cases, learners' needs are more complex or idiosyncratic, may not be sufficiently addressed by preexisting material, and pointing learners to videos they have already viewed could be frustrating, creating a negative affective state for learning [8]. In such cases, responsive human interaction may be a more useful resource for support [25]. In addition, by reducing the need and opportunity for human interaction, automated solutions risk exacerbating the lack of community in MOOCs, which has been a prominent issue associated with learning difficulties and dropouts [18].

The second category of approaches to addressing MOOC discussion forum disorder also begins with the automated analysis of discussion forum posts, but with the goal of supporting human interaction. These approaches have generally been either instructor-oriented or learner-oriented. Instructor-oriented approaches aim to assist instructors in making efficient and effective interventions in discussion forums. For example, [4] and [5] both aimed at optimizing instructor's intervention decisions with post recommendation systems. They built models based on instructor's intervention histories and tested how well the models could identify where instructors had chosen to intervene. While such models allow instructors to replicate their current intervention patterns more efficiently in the future, this approach overlooks the important question of if the existing patterns are desirable in the first place. Given that typically instructors may not have reviewed all existing posts before deciding where to intervene, and that subjectiveness and arbitrariness are common in these decisions [4], models based on prior intervention history alone may be insufficient, or even harmful, in meeting the goal of identifying posts in which intervention should occur.

Taking a different approach to support instructor intervention, [27] aimed to use existing discussion forum structures to inform instructors' reading decisions. They assumed that threads in the same sub-forums generally involve the same types of interactions and thus have common features that can help identify misplaced ones. Using the sub-forum titles (e.g. *Lectures*, *Assignments*, *Meetups*) as labels, they extracted five types of language-independent thread features: thread structure (e.g. length and breadth), underlying social network (e.g. number and density of users within a thread), popularity (e.g. number of views and votes), temporal dynamics (e.g. message rate), and content (e.g. quantity of text and hyperlinks). Using this feature set, the researchers built a model to identify misplaced threads in each sub-forum. Their results showed that non-linguistic features could be useful for identifying "small talk" threads (the majority of which were found in *Meetups*), but had limited utility in differentiating between other post categories.

Finally, [16] addressed instructor's intervention decision from a social network perspective. Instead of classifying and filtering posts based on their content, they aimed at identifying a small group of prominent learners to amplify the effect of (an inherently limited number of) instructor interventions. They modeled MOOC forums as a social network and developed an algorithm to identify the most influential learners. The underlying assumption was that by making responses to these learners' posts, the effects of the instructor's intervention would disseminate to a large number of other learners, though this logic was not tested in that study. If effective, this approach has the potential to broadcast instructor's influence most widely with minimum intervention cost; however it overlooks the specific learning needs of less prominent or networked learners. Additionally the most influential learners may not be the ones who most need the instructor's help.

In contrast to instructor-oriented approaches which have looked for general, high-impact ways to intervene, learner-oriented approaches have focused more on addressing specific learner's needs and promoting personalized interactions. For example, [34] aimed at assisting learners' reading choices and developed a model that recommends discussion threads that match learner's interests as reflected by their previous activities. Similarly, [33] built a question recommendation system that analyzes learner's intellectual and behavioral characteristics, and recommends people to answer questions considering relevance to their interests, level of intellectual challenge, and anticipated workload. These approaches provide learners with personalized interactions and make effective use of learner's participation efforts in discussion forums; they are thus one productive solution path to pursue. However, sorting posts based on historical learning behaviors creates a self-reinforcing narrowing of the field of vision that may not sufficiently reflect learners' evolving interests and needs. It also doesn't support them to explore for diversified learning opportunities. Thus modeling based only on prior activity and interests may not be desirable as the sole solution strategy.

Another approach that can help address learners' needs is topic-based post classification, which could help learners to browse a more organized set of posts for ones that may interest or benefit them. Topic-based classification has been an objective of several prior research efforts; however, the scope of the topics has not been clearly or consistently defined. [3] categorized posts as "course-related" (course-specific discussions and course logistics) versus "small talk" (for social purposes). They built a model based on topical features of these two categories to classify and rank posts according to their relevance to the course. While effective for its given purpose, their categorization did not distinguish

between discussions of the learning of course material and those about logistical and technical issues, which are substantially different. Such distinction was made in [30]'s framework for post topics in MOOC forums, but this initial study did not go beyond defining categories to create models to identify such posts. Thus principled identification of learning-content-related MOOC posts remains a promising, but as yet not fully realized, line of research.

3. THE CURRENT STUDY

In the current research, we extend the work of these prior studies to address information overload in MOOC discussion forums by providing a clear and theoretically justifiable way to categorize posts and build a classification model based on this categorization. Our approach is designed to address the needs of both instructors and learners in finding relevant posts by focusing on the identification of those which are substantively related to the content of the course material. While there are certainly other reasons that students and instructors may come to a discussion forum (for example to report or find out about technical/logistic problems or make social connections), we argue that communication about course content-related issues is (a) an important function of MOOC discussion forums; (b) where the greatest direct learning-related activity takes place; and (c) a particular kind of interaction, different in character from technical and social exchanges. Thus identification of content-related posts is the focus of this work.

From instructor's perspective, this can be helpful in directing their attention and domain expertise where it can be of most value. Other posts, such as those for social purposes, may not be directed at or need instructor intervention [27]. Moreover, posts that do require non-peer responses involve diversified issues, many of which would be best addressed by other members of the MOOC instructional team such as TA and technical staff [4]. Finally, even if an instructor chooses to engage with both content- and non-content-related posts, it may be most efficient to do in batches, focusing their cognitive efforts on similar tasks at a time.

From the learner's perspective, identifying threads based on whether they address questions about the course material is also desirable. While MOOC learners are diversified in learning goals, engagement patterns, and workload commitment, for the majority of learners who do more than just simply enroll, the purpose in taking a MOOC is learning-related, with logistic and social concerns serving supporting roles [19]. Thus, allowing learners to find the posts that address course-content-related issues can support learning. Importantly, by making the category *all* course-content-related posts, rather than those on specific topics as might be found with a search functionality, our approach allows learners to benefit from valuable peer questions they would not have thought to ask themselves. From the question-asking side, this approach also helps threads that substantially involve the course material get more attention and potential responses. In this way, the tool aims to foster learning interactions in the discussion forum.

3.1 Study Framing

The content of discussions in forums can be analyzed at various levels, such as thread, post, or sentence. When users navigate discussion forums, they usually browse the starting post in a thread to determine whether or not they are interested. Although a discussion thread may change direction as new participants join [30], the starting post reflects the primary intention of the thread initiator and largely scopes the content of the subsequent replies. Therefore, the starting post of a discussion thread is a useful unit of analysis for the purposes of this study.

In this work we define content-related starting posts as those that seek/provide help, opinions, or resources directly related to the course subject. This includes posts that ask or answer subject-related questions, share subject-related opinions, and those that share/comment on external resources related to course subject. Given the focus on identifying content-related starting posts, the rest of the threads are considered simply to be non-content-related, although we acknowledge that this broad framing could in the future be divided into subcategories such as socializing, technical problems, logistical questions etc. This binary classification goes beyond the work of [3], building on [30] to provide a refined criterion for distinguishing forum activities substantially related to learning of course material from other forum activities. To our knowledge, no previous work has focused in this way on identifying content-related discussion threads to assist human interaction in MOOCs.

Finally, we limit the scope of this initial work to MOOCs in the domain of statistics. Such limitation provides a useful test-bed for the viability of the approach in a local context, before addressing potentially more complicated questions about generalizability across domains. Statistics was chosen as the area of focus both because (like many MOOCs) it addresses a technical domain, and because several sets of relevant data for model development and testing were available. While a practical concern, accessibility to sufficient data from real MOOC discussions in particular kinds of courses is imperative for valid modeling and testing.

3.2 Research Issues

As content-related and non-content-related posts address substantially different topics, forum users may use language differently in such postings. Therefore, as a first step it is worth exploring whether or not content- and non-content-related posts in statistic MOOC discussion forums have distinct linguistic features that could be used for classification: **RQ 1**: *Do starting posts of content-related threads in a statistics MOOC discussion forum have distinguishing linguistic features from starting posts of non-content-related threads?*

If distinctive linguistic features are found, it may then be possible to use these to build a classifier that detects whether or not a post is content-related based on these: **RQ 2**: *Can linguistic features be used to create a model that reliably identifies starting posts of content-related threads in a statistics MOOC discussion forum?*

Apart from the linguistic features of starting posts, peer recommendations such as views and votes are potential indicators (or distractors) in the identification of content-related threads. These are important to test since they are common post-sorting features used by learners: **RQ 3**: *Are the number of views and votes useful in predicting content-related starting posts?*

If a reliable classifier is built, the next question is the degree to which it is useful on data generated in similar situations. The most similar situation in the case of MOOCs would be a new group of students taking the same course. The next degree of generalization would be to look at data from a new group of students taking a different course on the same general topic: **RQ 4**: *Does the model generalize to another offering of the same statistics MOOC and a different statistics MOOC?*

Finally, the proportion of content- and non-content-related posts in discussion forums may vary over time due to changes in learners' interests and needs as a MOOC proceeds [3]. For example, when a course just begins, learners may generate many posts to initiate study groups and ask for software-related technical assistance; when the course is about to conclude, learners may flock in to express gratitude and ask about credential issuance. Variation in the proportions of topics discussed may affect the performance of a linguistic model. For this reason, an important concern for the applicability of a model generated on completed course data to a live MOOC situation, is how well it performs on data from different time segments in a course: **RQ 5**: *How well does the model identify starting posts of content-related threads from particular time segments of a course?*

4. METHODS

4.1 Data Source

This study was conducted on data from three completed MOOCs offered in 2013 and 2014 on Stanford open-source platform Lagunita (initially called Stanford OpenEdX). The initial examination of linguistic features and modeling efforts were conducted on data from an offering of *Statistics in Medicine (StatMed'13)*. The generalizability of this model was tested on data from a later offering of the same course *(StatMed'14)*, and a different statistics course *Statistical Learning* (*StatLearn*). The usefulness of the number of views and votes was tested on all three courses; time segments could be evaluated only for the external test sets *StatMed'14* and *StatLearn*.

4.1.1 Course Contexts

Statistics in Medicine is an introductory course on probability and statistics with a special focus on statistics in medical studies. There was no prerequisite for taking the course. Optional modules in the course covered advanced math topics and basic data analysis in R. The course was 9 weeks long and the estimated weekly workload was 8 to 12 hours. Course materials included lecture videos and optional readings. Assessment consisted of quizzes, homework assignments, and a final exam. The course provided a discussion forum for interaction in nine topic areas (Figure 1.a). Learners were invited to post questions and comments about the course in the forums for response by other learners, the TAs and the instructor.

(a)

Discussion Topic Areas
Course Material Feedback
External Resources
General
Homework
Introductions
Platform Feedback
Study Group
Tech Support
Video

(b)

Discussion Topic Areas
Course Material Feedback
General
Platform Feedback
Quiz and Review
R and RStudio
Tech Support

Figure 1: Forum structures for (a) *StatMed* and (b) *StatLearn*.

Statistical Learning is an introductory-level course in supervised learning with a focus on regression and classification methods. The prerequisites included introductory courses in statistics, linear algebra, and computing. In this course, computing was done in R. Course materials included lecture videos and readings. The course was 9 weeks long with an estimated weekly workload of 3 hours. Assessment was based on the completion of quizzes. The course provided a discussion forum for interaction in six topic areas (Figure 1.b). Students were invited to post questions and comments about the course in the forum.

4.1.2 The MOOCPosts Dataset

The MOOCPosts Dataset [1] was obtained from researchers at Stanford University. The dataset consists of a selection of randomly chosen forum posts from several MOOC courses. The subset of the data pertaining to *StatMed'13*, *StatMed'14* and *StatLearn* was used in this study. Forum information provided in the dataset included the following: thread id; post id; post position in thread (starting post or reply post); post text; post creation date and time; number of times post was viewed; and number of votes post received. Thread titles were not included in the data set. A number of manual post annotations were also provided, but not used in this research. The number of threads and posts provided for each course are shown in Table 1.

Table 1: Number of threads and posts in the provided dataset

Course Name	# of Threads	# of Posts (starting posts & replies)
StatMed'13	844	3320
StatMed'14	310	1218
StatLearn	626	3030

4.2 Data Preparation

The entire set of starting posts for *StatMed'13* was coded in order to have sufficient data to train the model. Smaller sets of starting posts from *StatMed'14* and *StatLearn* were coded to serve as external test sets. This included all 310 of the starting posts from *StatMed'14*, and 300 randomly selected starting posts from *StatLearn*. Reply posts were used only for contextual information during manual coding when necessary.

Seven duplicated posts and one post containing foreign language were removed from the datasets. An additional seven posts were removed during the coding process because it was not possible to make a coding decision without the (missing) thread titles. After all post removals, 99% of data remained in the analysis, with the number of starting posts in *StatMed'13, StatMed'14* and *StatLearn* datasets being 837, 304 and 298 respectively. All anonymization codes were cleaned from post text before coding.

4.3 Coding

Each starting post was coded by two researchers as relating substantively to the course material or not according to the definition set out in Section 3.1. A coding guide with detailed category descriptions and examples was provided to both coders.

A rule of leniency towards the content-related category was adopted for borderline cases so as to maximally capture content-related linguistic features. Coder training was conducted on data not included in this study in cycles of 25 posts until reliability as indexed by Krippendorff's alpha was stable at an acceptable level

[1] http://datastage.stanford.edu/StanfordMoocPosts/.

Table 2: Top 30 features of content-related and non-content-related starting posts organized by category

Category	Content-related	Non-content-related
Course Subject	value, mean*, calculate, probability, p, difference*, standard, data, test	
Learning Process	understand, example, mean*, difference*, question*	
Question Words	how*, what*, which*, why*, does*, is*, are*	was*
Connectors	of, of_the, in, in_the, about, between, that, then, if, which,* why*, how*, what*	On
Existence / Condition	is*, are*, does*, not	was*, had*, have*
Course Tasks and Platform Resources	question*	final, homework, submit, the_final, answers, quizzes, exam, work*, download, videos, course, the_course, this_course
Pronouns	We	my, I_have, but_I, I_had, your, all*
Quality / Quantity		again, all*, time*
Effort / Action		work*, had*,have*, made
Appreciation		great, thank_you, thank

*An asterisk indicates a word that appears in more than one category due to different uses.

(α >.70). Study data was then coded in subsets of 100 to 169 posts for all courses. All differences were discussed and reconciled before the coders proceeded to the next subset. Coding reliability was good for all three courses: *StatMed'13* (α=.77); *StatMed'14*, (α=.82); *StatLearn,* (α*=.84).*

4.4 Feature Extraction and Modeling

Lightside Researcher's Workbench v2.3.1 was used to perform feature extraction on coded starting posts in *StatMed'13*, using the basic bag-of-words feature set and a rare threshold of 5. Unigrams and bigrams alone were most useful for characterizing and modeling posts. Stopwords were not removed from the feature lists because they were found to be helpful for modeling. This follows a general trend in the information retrieval systems to include stopwords in feature sets [22]. A total of 2410 features were extracted. After Arabic numbers, symbols, and features substantially incorporated by other features (e.g. "'m" is incorporated by "I_'m") were removed, 2236 features remained.

All 2236 features were then used to train a binary L2 regularized logistic regression classification model (thread starting post as content-related or not). The model was first evaluated by ten-fold cross-validation, and then on independent test sets from *StatMed'14* and *StatLearn*. Additional modeling was conducted to test the usefulness of the number of views and votes that starting posts received for classification. The predictive ability of the model for posts from different time segments in the course was tested by dividing the test datasets for *StatMed'14* and *StatLearn* respectively into three equal subsets according to posts' time of creation.

5. RESULTS

Content-related and non-content-related posts were relatively equal in proportion across all three courses with the percentage of content-related starting posts in *StatMed'13*, *StatMed'14,* and *StatLearn* being 47%, 54%, and 51% respectively.

5.1 Research Question 1: Linguistic Features of Thread Starting Posts

Stark differences were found between the top 30 features (ranked by kappa) of content- and non-content-related starting posts. At a basic level, the lists were composed of completely distinct terms. To probe this distinction further, features were organized into categories based on examination of their uses in the text (Table 2). The vast majority (85%) of the features across both categories did not appear to be specific to the course's subject domain. Many features of content-related starting posts were associated with the process of learning, involved question words or terms describing relationships between ideas (connectors). For content-related starting posts only, several linguistic features related to the course domain were found. The features of non-content-related starting posts did not appear to be related to the course domain. Rather they were terms related to course tasks, resources, and those that refer to the course itself. In addition, several terms referring to effort/action, quality/quantity, appreciation, and pronouns were found in this category. Both categories contained several terms related to existence and condition.

Within categories common to both content- and non-content-related features, stark differences were also observed. The only content-related feature in the Course Task and Platform category ("question") refers to a more specific form of course task than the six non-content-related features (e.g. "homework", "exam"). Similarly, the category of Pronouns was dominated by non-content-related features containing a variety of first-person *singular* references (e. g. "I") while the only content-related Pronoun feature is the first person *plural* pronoun "we." Existence /Condition was the only feature category that did not show stark difference between content- and non-content-related features.

Collectively these findings indicate content-related starting posts in *StatMed'13* have distinguishing linguistic features from non-content-related ones, irrespective of specific course vocabulary.

5.2 Research Question 2: Identifying Content-Related Threads

The model created showed reasonably good reliability in identifying content-related starting posts in *StatMed'13* (accuracy=0.80, kappa=0.61). Recall was 79%: 315 out of 397 content-related posts were identified. Precision was 79%: an additional 82 posts were (incorrectly) classified as content-related, leading to a total of 397 that were assigned this label[2]. These results show that, at a basic level, linguistic modeling can reliably identify starting posts of content-related threads in a statistics MOOC discussion forum.

5.3 Research Question 3: Predictive Value of the Number of Views and Votes

Adding the number of views and votes each starting post received as additional features to the base model did not produce substantial improvement (Table 3). In addition, models created using only these features produced very poor classification results (all kappa < .13). These results indicate the number of views and votes are not useful for identifying content-related starting posts.

Table 3: Reliability statistics of the base model and with additional features for *StatMed'13* (cross-validation)

	Base model	+ #of views	+ #of votes
Accuracy	0.80	0.80	0.80
Kappa	0.61	0.60	0.61
Recall	0.79	0.79	0.78
Precision	0.79	0.79	0.80

5.4 Research Question 4: Testing the Model's Cross-Course Generalizability

To examine the generalizability of the model, it was first tested on the data from a second offering of the same course, *StatMed'14*. Results were consistent with those from the initial data-set (accuracy=0.81, kappa=0.62). Recall was 85%: 140 out of 165 content-related starting posts were identified. Precision was 81%: an additional 32 posts were (incorrectly) classified as content-related, leading to a total of 172 that were assigned this label. Similar to *StatMed'13*, adding the number of views ad votes did not improve the model (see Table 4). As *StatMed'14* is a second offering of *StatMed'13*, it's not surprising that the posts have highly similar contents and thus similar linguistic features. While these results are promising, we need to go beyond multiple offerings of the same class to look at if the model can work across different courses in the same subject area.

To find out how well the model performs on a different course on a similar subject, the model was tested on the coded starting posts from *StatLearn*. *StatLearn* was an advanced course for learners with introductory knowledge of statistics. Although *StatMed'13* and *StatLearn* both contained math and statistics content, they were independent courses. Nonetheless the results were again largely consistent with those of the previous trials (accuracy=0.80, kappa=0.60). Recall was 90%: 137 out of 153 content-related posts were identified. Precision was 76%: an additional 44 posts were (incorrectly) classified as content-related, leading to a total of 181 that were assigned this label. These results show that the model has reasonably good generalizability to a different course on a similar subject. Similar to the two *StatMed* courses, adding the number of views and votes did not improve the model (see Table 5).

Table 4: Reliability statistics of the base model and with additional features for *StatMed'14* (test set)

	Base model	+ #of views	+ #of votes
Accuracy	0.81	0.81	0.81
Kappa	0.62	0.61	0.61
Recall	0.85	0.84	0.86
Precision	0.81	0.81	0.80

Table 5: Reliability statistics of the base model and with additional features for *StatLearn* (test set)

	Base model	+ #of views	+ #of votes
Accuracy	0.80	0.80	0.82
Kappa	0.60	0.59	0.63
Recall	0.90	0.90	0.90
Precision	0.76	0.75	0.78

5.5 Research Question 5: Testing the Model on Different Time Segments in a Course

The proportion of content-related starting posts in different time segments varied in both *StatMed'14* and *StatLearn* (Table 6). In *StatLearn,* the distribution pattern was similar to that anticipated: a higher concentration of content-related starting posts in the middle of the course. In *StatMed'14,* however, the proportion of content-related starting posts increased continually over the time segments. This may be related to the fact that *StatMed'14* had a final exam at the end of the course while *StatLearn* did not.

Table 6: Results of *StatMed'13* model on data from different time segments in *StatMed'14* and *StatLearn*

Course	Dataset	% C	A	K	R	P
StatMed'14	1st third (n=102)	47%	0.83	0.67	0.92	0.77
	2nd third (n=101)	53%	0.76	0.53	0.72	0.81
	3rd third (n=101)	62%	0.84	0.66	0.90	0.85
StatLearn	1st third (n=100)	38%	0.75	0.50	0.84	0.63
	2nd third (n=99)	65%	0.77	0.46	0.89	0.78
	3rd third (n=99)	52%	0.88	0.76	0.94	0.84

% C =percent of content-related starting posts in data subset, A=accuracy, K=kappa, R=recall, P=precision

To test how well the model classified posts from different time segments in the courses, the model was applied to each of three temporal subsets in *StatMed'14* and *StatLearn* (Table 6). The

[2] It is coincidence that the number of posts coded as content-related and the number identified by the model were equal.

model achieved consistently good classification results (accuracy > 0.75, kappa > 0.46) across time segments. Overall model reliability as indexed by kappa was lowest for the middle time segments in both courses. For *StatMed'14* this seemed to be related to a reduced level of recall. For *StatLearn*, both recall and precision remained high for the middle time segment, thus the depression in kappa is due to the greater proportion of actual content-related posts in the subset[3]. The first time segment of *StatLearn* also showed a substantially lower level of precision than all other segments across the two courses (0.63 compared with 0.77 and higher). This may be related to the low level of content-related posts in this segment.

6. DISCUSSION

This research investigated the automatic identification of content-related threads as an approach to address problems of information overload and chaos in MOOC discussion forums. Building on prior findings that MOOC discussion forum posts can be categorized as pertaining to the learning of course material or not [30] and that classification tools can sort posts for users based on relevance [3], this study extended the line of research by building a classification model to support students and teachers in finding content-related posts. Results showed that content-related starting posts in a statistics MOOC had distinguishing linguistic features from non-content-related ones that could be used reliably for identification. Highly weighted extracted features were mainly terms that did not appear to relate to the topic of statistics; the number of views and votes a post received were not useful features for the classification task. The model generated demonstrated good generalizability to both a second offering of the same course and a different course on statistics; classification was also successful across different time segments of the courses.

6.1 Notable Features of the Model

Several findings in our study diverge from previous work. First, the success of our model upholds the value of simple modeling. Literature shows that many models built for classification purposes in MOOC discussion forums have been based on complicated feature sets [27, 3], but our model was built using only unigram and bigram linguistic features and achieved consistently good classification results across three course (accuracy>.80, kappa>.59, recall>.79, precision>.78). This indicates that complex feature combinations are not necessarily needed for useful classification results.

Second, our findings affirm the indicative power of linguistic features for topic-based post classification in MOOC discussion forums. Notably, in contrast to the customary practice of excluding stopwords when choosing linguistic feature for classification modeling of MOOC posts [1, 4], our test on feature combinations found that including stopwords improved the model. In fact, a large proportion of the top features of content- and non-content-related starting posts were commonly used stopwords. Therefore it is worthwhile to further explore and interpret the usefulness of stopwords in topic-based classification.

[3] This exemplifies the critique that has been levied against the kappa statistic as being overly sensitive to the base rate proportions, resulting in a low value even when the actual ratings themselves have high levels of accuracy.

Finally, our findings that the number of views and votes were not helpful for our classification purposes suggest the need to reexamine the role of these features in MOOC discussion forums. Our results echo previous findings [6] that these features are not good indicators of the content-relatedness of MOOC discussion posts. In prior work, [27] found the number of views and votes to be useful in detecting social "small talk" threads, but not for distinguishing between other kinds of posts as we sought to do here. Notably this is different from the situation in more formal education contexts. For example, in a traditional discussion forum in a university class the content of messages which were "liked" (equivalent to a vote) were of higher cognitive complexity than those messages which were not liked [21]. The power of learning context difference is very apparent here: while [21] studied a small graduate-level course which emphasized online discussions as a site for social knowledge construction, MOOC discussion forums are designed to support more diversified functions including learning support, course management, technical support, feedbacks, and social interaction. Given such varied purposes of use, the indicative power of the number of views and all-purpose votes are questionable. Instead, it may be fruitful to have a (limited) variety of types of votes available to users; for example critical content issue and pressing logistical / technical concern. Such tailored votes could be useful indicators in future prediction models.

6.2 Practical Benefits

The model built using the extracted features had not only good internal validity, but also good generalizability on a second offering of the same course. This is of particular practical value, as given the high cost of development, many MOOC instructors plan to offer the same course multiple times [15]. Moreover, the model showed almost equivalent reliability on another independent statistics MOOC, indicating that within this topic area, there seems to be linguistic consistencies in the way questions are asked across different courses. These findings suggest that the developed model could be applicable to a variety of courses on similar topics offered by different institutions. Importantly, the verification that the model works for different time segments in MOOCs suggests substantial value for real-time application.

The model can be applied to MOOC teaching and learning in multiple ways. The most straightforward approach is to use it to drive a live thread-sorting feature in the interface that can help instructors and learners to navigate the discussion forums and find threads related to the learning of course material more efficiently. This can inform instructors' intervention decisions in a live course situation and help learners find learning interaction opportunities more easily. To give a sense of how this would change instructors' and learners' experiences using the forums, consider that the proportion of content-related posts in *StatMed'13* was 47%. If an instructor or learner read all discussion threads indiscriminately, more than half of their effort would be consumed by ones not related to the course content. With the help of a filter based on our model, users could narrow their task to less than half the total threads in the forum (397 out of 837 starting posts were labeled as content-related) of which almost 80% would be content-related ones. In courses where the actual proportion of content-related posts is lower (for example [6]) use of the model would bring even larger benefits.

Another use of the model could be a live tagging tool that analyzes the posts being composed and prompts learners to tag

them as content-related or not, so as to make them easier to locate. Taken a step further, the model could be used to identify learners who have particular posting patterns, for example asking some number of content-related questions or asking a large volume of exclusively non-content-related questions. Such patterns may be indicative of particular goal-orientations (e.g. performance, mastery, work-avoidance) which are more or less beneficial for learning [32]. This creates the potential for targeted interventions that could help encourage learners to adjust their engagement patterns. However, whether it is appropriate to adopt such a proactive learning-oriented intervention will be largely dependent on an instructor's perspectives on learning and instruction and the overarching goals of the course.

Moreover, the model could also be applied as a collector that helps instructors and researchers gather content-related posts from concluded courses. As content-related posts contain interactions related to the learning of course material, these posts are quality data for analysis aiming at improving course design and understanding learning in MOOCs.

6.3 Potential Domain-Generality

Findings showed that content-related and non-content-related starting posts in a statistics MOOC had distinct linguistic features. The top features of content-related starting posts were terms related to the process of learning and understanding, question words, terms that connect ideas, and the course topic of statistics while the top features of non-content-related starting posts were terms related to the course tasks and platform, effort / action, appreciation and pronouns. Notably the vast majority of features across *both* categories were language related to the course-taking process (learning, asking, connecting, using the platform, fulfilling tasks) or interpersonal concerns (pronouns, appreciation) rather than topic-specific language about statistics. While somewhat counterintuitive, the contradiction of the notion that common linguistic features are vulnerable to domain influence [27] is actually quite logical. Although content-related posts may contain more statistical language *generally*, given the diversity of statistics topics about which questions might be asked, there are few *particular* statistics words that will be used frequently enough to be identified as a top feature. Thus it is the language that surrounds the asking of such questions (interrogatives, learning process words, connectors) that are used again and again regardless of the specific focus of the question being asked.

Though not surprising, the predominance of seemingly domain-unrelated features is powerful with respect to the potential for a linguistic model of content-relatedness to generalize across courses and domains of learning. This runs in contrast to the recent result by [11] who found that generalized analytics built to predict academic success (course completion and grade) across multiple courses both had less predictive power than those built for specific courses and consistently misidentified the most relevant predictors in a given situation. They concluded that attention to the instructional conditions and the ways technology is used in particular courses must be considered in deciding what predictive factors to include and how to model them. [11] raises an important concern for the learning analytics community: the tension between a desire for portable models that give leverage to our work and the need to take into account the particularities of individual learning contexts. Acknowledging this concern, we limit claim of the potential generality of our model for MOOC discussions to the condition that it is being uses to classify starting posts in forums designed for the purpose of Q&A[4]. As we have discussed elsewhere [6], there are a variety of other uses to which discussion forums might be put. The similarity between the linguistic patterns generated in such situations and here cannot be assumed and thus would need to be investigated empirically. In addition, within a Q&A format, even if the key words used are not domain-specific, there may be differences in the ways in which learners in different domains ask questions (for example compare the non-domain specific words in the following questions: "Is it possible to write this algebra function with one definition?" versus "I am wondering about if, when and how literature can inspire people to violence"). We are currently exploring the domain generality of the model generated on the statistics discussions by applying it to increasingly distant subject areas.

6.4 Limitations and Future Research

This study has limitations with respect to the conceptualization, coding and modeling of "content-relatedness", and the use of thread starting posts as the target for modeling.

First, the choice to classify based on content-relatedness was made based on common concerns of educators from a learning perspective. Posts on other topics, such as course logistics may be important to instructors in other ways and can be targeted accordingly. In defining the scope of "content-related" there was a challenge in being inclusive enough to maximally meet the practical needs of forum users and at the same time control the level of noise in sorting results. Thus the delineation of the content/non-content boundary required decisions of inclusion and exclusion that could vary depending on a forum user' perspective. The categorization in this research was developed based on negotiations among the three researchers, all experienced educators and online teachers. For example, posts that discussed external resources related to the course material were considered content-related in that such comments may reflect important ideas or misconceptions about the course content and thus are worthy of instructors' attention. Similarly, posts that shared resources related to the course material were also considered content-related as they could provide learners with extended learning opportunities. However, in reality, instructors and learners may want to target a more narrow conception of content-relatedness due to the limited time and energy they have available.

Second, there were several practical challenges in labeling posts as content- versus non-content-related. The most notable of these occurred with error-reporting posts (in which students reported aspects of the course materials that they thought were in error). Error-reporting posts were difficult to code for several reasons. First, there were several types of possible errors. Some were low-level, such as typos and speaking errors in lecture videos. These were ostensibly related to the course content but conceptually were unlikely to lead to meaningful learning. Other posts reported apparent higher-level errors, such as misapplication or inaccurate interpretation of concepts, but these posts were not always correct in what they perceived as an error. These posts revealing a

[4] It could be argued that generalizability should also be limited to the specific discussion tool used (EdX Forum), however it is not clear that the structure of this tool influences the language used within it. In addition as noted in [23], the majority of discussion forum software currently in use employ a very similar form.

learner's misunderstanding are critical ones for the instructor (or other peers) to address. Finally some posts (correctly or incorrectly) reported errors related to course logistics, policy, or technical issues, but used similar "error reporting" language to the other error types. Moreover, learners reported perceived errors in diverse linguistic styles. Some were assertive and straightforward while others conveyed confusion and/or suspicion. This did not seem to be related to whether what they were reporting was actually an error or not. Given all these factors, making coding decisions for these posts demanded methodological skills, course subject knowledge, and understanding of specific course and discourse context, and in many cases had to be made based on checking available reply posts for more context. The difficulties and ambiguity surrounding such posts for human raters raise questions about their quality as a benchmark for machine classification. In future work a separate "error reporting" category would worth exploring.

Third, similar challenges arose from the discovery that learners often had more than one purpose in making a post, combining content-related and non-content-related concerns. Since coding criteria was based on the presence (or absence) of content-related comments, such posts were not hard for manual coding. However, the posts themselves contained mixed linguistic features which could cause confusion in the training and testing of the model. It is thus not surprising that many of the misclassified posts were ones which contained multiple (content- and non-content-related) topics. This is a more difficult problem to address since it is impractical (though attractive) to oblige students to address one issue per post. One potential solution could be to manually segment a data corpus of messages a posteriori (for instance at the sentence level) for initial model training and subsequently use this model to detect and flag the presence of both content- and non-content-related topics in a post or to calculate the percentage of content-relatedness in each post.

Finally, the model was built with starting post as the analysis unit, but it is possible for threads that start with a non-content related post to evolve into a content-related discussion (or visa-versa). Thus, the current approach could miss emergent content-related reply streams or incorrectly identify threads which diverge from an initial content-related focus. Future research can examine the frequency of such occurrences and, if warranted, include reply posts into the data to provide more nuanced diagnostics about the discussions so as to better detect content-related portions of discussions within threads. Alternatively, it is also possible to use features extracted from replies to content- and non-content-related starting posts as additional indicators to assist in the categorization of threads overall.

7. CONCLUSION

Due to the quantity and disorganization of posts in MOOC discussion forums, learners and instructors often face an overwhelming workload to find the relevant ones to read and which to reply [3, 15]. Specifically, a central goal of MOOCs is learning, yet a relatively small percentage of posts are substantially related to understanding the subject matter of the course. This study targeted the automatic identification of such content-related posts as a potential tool to help students and instructors navigate MOOC discussion forums more effectively. The resulting model, which reliably used linguistic features to identify content-related starting posts across several statistics courses and time periods within these courses, makes both theoretical and practical contributions. Theoretically, the study enhances understanding of the linguistic qualities of discussion posts associated with content-relatedness. It also raises doubts about the use of number of views and votes in MOOC forums as universal indicators of posts quality. Practically, it is a step toward a tool for using this information to help learners and instructors more effectively find content-related posts in MOOCs and thus derive the intended learning benefits these online discussions have the potential to provide.

8. ACKNOWLEDGMENTS

We thank Stanford University and the MOOCPosts Dataset team for their assistance in accessing and working with the data. We also acknowledge Wan Qi Jin for her helpful discussions with us about the ideas expressed in this work.

9. REFERENCES

[1] Agrawal, A., Venkatraman, J., Leonard, S., and Paepcke, A. 2015. YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. In *Proceedings of the 8th International Conference on Education Data Mining* (Madrid, Spain, June 26 - 29, 2015). ACM, New York, NY, USA, 297-304.

[2] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. T. 2013. Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice In Assessment,* 8, 13-25.

[3] Brinton, C.G., Chiang, M., Jain, S., Lam, H., Liu, Z., and Wong, F.M.F. 2014. Learning about social learning in MOOCs: from statistical analysis to generative model. *IEEE Transactions on Learning Technologies,* 7, 4, 346-359. DOI= 10.1109/TLT.2014.2337900.

[4] Chandrasekaran, M. K., Kan, M. Y., Tan, B. C., and Ragupathi, K. 2015. Learning instructor intervention from MOOC forums: early results and issues. In *Proceedings of the 8th International Conference on Education Data Mining* (Madrid, Spain, June 26-29, 2015). ACM, New York, NY, USA, 218-225.

[5] Chaturvedi, S., Goldwasser, D., and Daumé III, H. 2014. Predicting instructor's intervention in MOOC forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, USA, June 23 - 25, 2014). ACL, Baltimore, Maryland, USA, 1501-1511.

[6] Cui, Y., and Wise, A. F. 2015. Identifying content-related threads in MOOC discussion forums. In *Proceedings of the 2nd ACM Conference on Learning @ Scale* (Vancouver, Canada, March 14 - 18, 2015). ACM, New York, NY, USA, 299-303. DOI= 10.1145/2724660.2728679.

[7] DeBoer, J., Ho, A.D., Stump, G.S. and Breslow, L. 2014. Changing "course": reconceptualizing educational variables for MOOCs. *Educational Researcher, 43*, 2 (March. 2014), 74-84. DOI= 10.3102/0013189X14523038.

[8] D'Mello, S., and Graesser, A. 2012. Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 2 (April. 2012), 145-157. DOI= 10.1016/j.learninstruc.2011.10.001.

[9] Dringus, L. P., and Ellis, T. 2005. Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45,* 141-160.

[10] Gaebel, M. 2013. *EUA occational papers: Massive Open Online Courses*. White Paper. European University Association, Brussel, Belgium.

[11] Gašević, D. Dawson, S., Rogers, T. and Gašević, D. 2015. Learning analytics should not promote one size fits all: the effects of instructional conditions in predicating academic success. *The Internet and Higher Education*, DOI= 10.1016/j.iheduc.2015.10.002

[12] Gütl, C., Rizzardini, R. H., Chang, V., and Morales, M. 2014. Attrition in MOOC: lessons learned from drop-out students. *Learning Technology for Education in Cloud. MOOC and Big Data Communications in Computer and Information Science* 446 (2014), 37-48. DOI= 10.1007/978-3-319-10671-7_4.

[13] Herring, S. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication,* 4, 4. DOI= 10.1111/j.1083-6101.1999.tb00106.x.

[14] Huang, J., Dasgupta, A., Ghosh, A., Manning, J., and Sanders, M. 2014. Superposter behavior in MOOC forums. In *Proceedings of the 1st ACM conference on Learning @ scale* (Atlanta, USA. March 4 - 5, 2014). ACM, New York, NY, USA, 117-126. DOI= 10.1145/2556325.2566249.

[15] Hollands, F. M., and Tirthali, D. 2014. *MOOCs:expectations and reality*. Report. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

[16] Jiang, Z., Zhang, Y., Liu, C., and Li, X. 2015. Influence analysis by heterogeneous network in MOOC forums: what can we discover? In *Proceedings of the 8th International Conference on Education Data Mining* (Madrid, Spain, June 26 - 29, 2015). ACM, New York, NY, USA, 242-249.

[17] Khalil, H. and Ebner, M. 2013. "How satisfied are you with your MOOC?" - a research study on interaction in huge online courses. In *Proceedings of EdMedia 2013* (Victoria, Canada, June 24, 2013). AACE. 830-839.

[18] Khalil, H. and Ebner, M. 2014. MOOCs completion rates and possible methods to improve retention - a literature review. In *Proceedings of EdMedia 2014* (Tampere, Finland, June 23 2014). AACE 1305-1313.

[19] Kizilcec, R. F., Piech, C., and Schneider, E. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (Leuven, Belgium, 8 - 12 April, 2013). ACM New York, NY, USA, 170-179. DOI= 0.1145/2460296.2460330.

[20] Lerman, K., and Hogg, T. 2014. Leveraging position bias to improve peer recommendation. *PLoS ONE,* 9, 6. DOI= 10.1371/journal.pone.0098914.

[21] Makos, A., Lee, K., & Zingaro, D. 2014. Examining the characteristics of student postings that are liked and linked in a CSCL environment. *British Journal of Educational Technology*, 46, 6, 1281-1294.

[22] Manning, C. D., Raghavan, P., and Schütze, H. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

[23] Marbouti, F. and Wise, A. F. 2015. *Starburst*: a new graphical interface to support productive engagement with others' posts in online discussions. *Educational Technology Research & Development*, 1-27. DOI= 10.1007/s11423-015-9400-y.

[24] McGuire, R. 2013. Building a sense of community in MOOCs. *Campus Technology*, 26, 12, 31-33. Retrieved October 11, 2015, from https://campustechnology.com/articles/2013/09/03/building-a-sense-of-community-in-moocs.aspx.

[25] Moore, M.G. 1989. Editorial: three types of interaction. *American Journal of Distance Education*, 3, 2, 1-7. DOI= 10.1080/08923648909526659.

[26] Peters, V. L., and Hewitt, J. 2010. An investigation of student practices in asynchronous computer conferencing courses. *Computers & Education,* 54, 951-961. DOI= 10.1016/j.compedu.2009.09.030.

[27] Rossi, L.A and Gnawali, O. 2014. Language independent analysis and classification of discussion threads in Coursera MOOC forums. In *Proceedings of 2014 IEEE 15th International Conference on Information Reuse and Integration* (San Francisco, USA, August 13 - 14, 2014). IEEE, 654-661. DOI= 10.1109/IRI.2014.7051952.

[28] Shah, D. 2014. MOOCs in 2014: breaking down the numbers. *EdSurge News.* Retrieved March 3, 2015, from https://www.edsurge.com/n/2014-12-26-moocs-in-2014-breaking-down-the-numbers.

[29] Stephens-Martinez, K., Hearst, M. A., and Fox, A. 2014. Monitoring MOOCs: which information sources do instructors value?. In *Proceedings of the 1st ACM Conference on Learning@ Scale* (Atlanta, USA. March 4 - 5, 2014). ACM, New York, NY, USA, 79-88. DOI= 10.1145/2556325.2566246.

[30] Stump, G. S., DeBoer, J., Whittinghill, J., and Breslow, L. 2013. Development of a framework to classify MOOC discussion forum posts: methodology and challenges. In *Proceedings of NIPS 2013 Workshop on Data Driven Education* (Lake Tahoe, United States, December 5 - 8, 2013). NIPS Foundation, 1-20.

[31] Trentin, G. 2000. The quality-interactivity relationship in distance education. *Educational Technology,* 40, 1, 17-27.

[32] Wise, A., Marbouti, F., Hsiao, Y. and Hausknecht, S. 2012. A survey of factors contributing to learners' "listening" behaviors in asynchronous online discussions. *Journal of Educational Computing Research*, 47, 4, 461-480.

[33] Yang, D., Adamson, D., and Rosé, C. 2014. Question recommendation with constraints for massive open online courses. In *Proceedings of the 8th ACM Conference on Recommender* systems (Foster City, USA, October 6 -10, 2014). ACM, New York, NY, USA. 49-56. DOI= 10.1145/2645710.2645748.

[34] Yang, D., Piergallini, M., Howley, I., and Rose, C. 2014. Forum thread recommendation for massive open online courses. In *Proceedings of the 7th International Conference on Educational Data Mining* (London, UK, July 4 - 7, 2014). ACM, New York, NY, USA, 257-260