Learning Instructor Intervention from MOOC Forums: Early Results and Issues

Muthu Kumar
Chandrasekaran

Min-Yen Kan

Bernard C.Y. Tan
Kiruthika Ragupathi

1 Web IR / NLP Group (WING)
2 Department of Information Systems
3 Centre for Development of Teaching and Learning
National University of Singapore

{muthu.chandra, kanmy}@comp.nus.edu.sg, {pvotcy, kiruthika}@nus.edu.sg

ABSTRACT
With large student enrollment, MOOC instructors face the unique challenge in deciding when to intervene in forum discussions with their limited bandwidth. We study this problem of instructor intervention. Using a large sample of forum data culled from 61 courses, we design a binary classifier to predict whether an instructor should intervene in a discussion thread or not. By incorporating novel information about a forum's type into the classification process, we improve significantly over the previous state-of-the-art.

We show how difficult this decision problem is in the real world by validating against indicative human judgment, and empirically show the problem's sensitivity to instructors' intervention preferences. We conclude this paper with our take on the future research issues in intervention.

Keywords
MOOC; Massive Open Online Course; Instructor Intervention; Discussion Forum; Thread Recommendation

Categories and Subject Descriptors
H.3.3. [Information Search and Retrieval]: Information filtering; K.3.1. [Computers and Education]: Computer Uses in Education

1. INTRODUCTION
MOOCs scale up their class size by eliminating synchronous teaching and the need for students and instructors to be co-located. Yet, the very characteristics that enable scalability of massive open online courses (MOOCs) also bring significant challenge to its teach-

2 This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.
ing, development and management [7]. In particular, scaling makes it difficult for instructors to interact with the many students — the lack of interaction and feelings of isolation have been attributed as reasons for why enrolled students drop from MOOCs [9].

MOOC discussion forums are the most prominent, visible artifact that students use to achieve this interactivity. Due to scale of contributions, these forums teem with requests, clarifications and social chatting that can be overwhelming to both instructors and students alike. In particular, we focus on how to best utilize instructor bandwidth: with a limited amount of time, which threads in a course’s discussion forum merit instructor intervention? When utilized effectively, such intervention can clarify lecture and assignment content for a maximal number of students, promoting the enhancing the learning outcomes for course students.

To this end, we build upon previous work and train a binary classifier to predict whether a forum discussion thread merits instructor intervention or not. A key contribution of our work is to demonstrate that prior knowledge about forum type enhances this prediction task. Knowledge of the enclosing forum type (i.e., discussion on lecture, examination, homework, etc.) improves performance by 2.43%; and when coupled with other known features disclosed in prior work, results in an overall, statistically significant 9.21% prediction improvement. Additionally, we show that it is difficult for humans to predict the actual interventions (the gold standard) through an indicative manual annotation study.

We believe that optimizing instructor intervention is an important issue to tackle in scaling up MOOCs. A second contribution of our work is to describe several issues pertinent for furthering research on this topic that emerge from a detailed analysis of our results. In particular, we describe how our work at scale details how personalized and individualized instructor intervention is — and how a framework for research on this topic may address this complicating factor through the consideration of normalization, instructor roles, and temporal analysis.

2. RELATED WORK
While the question of necessity of instructor’s intervention in online learning and MOOCs is being investigated [12, 20], technologies to enable timely and appropriate intervention are also required.
The pedagogy community has recognized the importance of instructor intervention in online learning prior to the MOOC era (e.g., [10]). Taking into consideration the pedagogical rationale for effective intervention, they also proposed strategic instructor postings: to guide discussions, to wrap-up the discussion by responding to unanswered questions, with “Socrates-style” follow-up questions to stimulate further discussions, or with a mixture of questions and answers [13]. However, these strategies must be revisited when being applied to the scale of typical MOOC class sizes.

Among works on forum information retrieval, we focus on those that focus on forum moderation as their purpose is similar to the instructor’s role in a course forum. While early work focused on automated spam filtering, recent works shifted focus towards curating large volumes of posts on social media platforms [4] to distil the relevant few. Specifically, those that strive to identify thread solvedness [21, 8] and completeness [3] are similar to our problem.

Yet all these work on general forums (e.g., troubleshooting, or threaded social media posts) are different from MOOC forums. This is due to important differences in the objectives of MOOC forums. A typical thread on a troubleshooting forum such as Stack Overflow is centers on questions and answers to a particular problem reported by a user; likewise, a social media thread disseminates information mainly to attract attention. In contrast, MOOC forums are primarily oriented towards learning, and also aim to foster learning communities among students who may or may not be connected offline.

Further, strategies for thread recommendation for students such as [23] may not apply in recommending for instructors. This difference is partially due to scale: while the number of students and threads are large, there are few instructors per course. In this case, reliance on collaborative filtering using a user–item matrix is not effective. Learning from previous human moderation decisions [2], therefore, becomes the most feasible approach. Prior work on MOOC forums propose categorisation of posts [16, 5, 19] to help instructors identify threads to intervene. Chaturvedi et al. [5], the closest related work to ours, show each of the four states of their sequence models to predict instructor intervention to be distributed over four post categories they infer. In this paper, we use their results for comparison.

Different from previous works, we propose thread–level categories rather than post–level categories, since an instructor needs to first decide on a thread of interest. Then they need to read its content, at least in part, before deciding whether to intervene or not. We make the key observation that show thread–level categories identified as by the forum type, help to predict intervention.

Previous work has evaluated only with a limited number of MOOC instances. One important open question is whether those reported results represent the diverse population of MOOCs being taught. In this paper, we address this by testing on a large and diverse cross-section of Coursera MOOC instances.

3. METHODS

We seek to train a binary classifier to predict whether a MOOC forum thread requires instructor intervention. Given a dataset where instructor participation is labeled, we wish to learn a model of thread intervention based on qualities (i.e., features) drawn from the dataset. We describe our dataset, the features distilled used for our classifier, how we obtain class labels, and our procedure for instance weighting in the following.

### Table 1: Thread statistics from our 61 MOOC Coursera dataset and the subset of 14 MOOCs, used in the majority of our experiments.

<table>
<thead>
<tr>
<th>Forum type</th>
<th>All</th>
<th>Intervened</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># threads</td>
<td># posts</td>
</tr>
<tr>
<td>D61 Corpus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homework</td>
<td>14,875</td>
<td>127,827</td>
</tr>
<tr>
<td>Lecture</td>
<td>9,135</td>
<td>64,906</td>
</tr>
<tr>
<td>Errata</td>
<td>1,811</td>
<td>6,817</td>
</tr>
<tr>
<td>Exam</td>
<td>822</td>
<td>6,285</td>
</tr>
<tr>
<td>Total</td>
<td>26,643</td>
<td>205,835</td>
</tr>
<tr>
<td>D14 Corpus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homework</td>
<td>3,868</td>
<td>31,255</td>
</tr>
<tr>
<td>Lecture</td>
<td>2,392</td>
<td>13,185</td>
</tr>
<tr>
<td>Errata</td>
<td>326</td>
<td>1,045</td>
</tr>
<tr>
<td>Exam</td>
<td>822</td>
<td>6,285</td>
</tr>
<tr>
<td>Total</td>
<td>7,408</td>
<td>51,770</td>
</tr>
</tbody>
</table>

Figure 1: Typical top-level forum structure of a Coursera MOOC, with several forums. The number of forums and their labels can vary per course.

3.1 Dataset

For our work, we collected a large-scale, multi-purpose dataset of discussion forums from MOOCs. An important desideratum was to collect a wide variety of different types of courses, spanning the full breadth of disciplines: sciences, humanities and engineering. We collected the forum threads from 61 completed courses from the Coursera platform, from April to August 2014, amounting to roughly 8% of the full complement of courses that Coursera offers.

For each course, we first assigned each forum to one of several types based on the forum’s title. For this study we focus on threads that originated from four prevalent types: (i) errata or course material errors, (ii) video lectures, (iii) homework, assignments or prob-

---

1. We collected all threads and their component posts from four subforum categories as in Section 3.1. We did so, as we hypothesize that they would necessitate different levels of instructor intervention and that such interventions may be signaled by different features.

2. The full list of courses is omitted here due to lack of space.

3. As of December 2014, Coursera, a commercial MOOC platform: [https://www.coursera.org](https://www.coursera.org), hosted 761 courses in English spanning 25 different subject areas.

well as any subsequent posts in a thread need to be removed. We
also do not use the number of votes or views in a thread as these
are summary counts that are influenced by intervention

We used regular expressions to further filter and canonicalize cer-
tain language features in the forum content. We replaced all mat-
ematical equations by $<MATH>$, URLs by $<URLREF>$ and ref-
ences to time points in lecture videos by $<TIMEREF>$. We re-
oved stopwords, case-folded all tokens to lowercase, and then
indexed the remaining terms and computed the product of term
frequency and inverse thread frequencies ($tf \times itf$) for term im-
portance. The weighted terms form a term vector that we further
ormalized using the L2–norm. Other real-valued features were
max–min normalized. Categorical features such as the forum type
were encoded as bit vectors.

Each thread is represented as bag of features consisting of terms
and specific thread metadata as disclosed below. We indicate each
new feature that our study introduces with an asterisk.

1. Terms (unigrams);
2*. Forum type to which the thread belongs: Figure 3 shows a
clear difference in intervention ratio, the ratio of number of
threads intervened to those that weren’t, across different fo-
rum types. Forum type thus emerges as a feature to use to
discriminate threads worthy of intervention. The forum type
encapsulating the thread could be one of homework, lecture,
exam or errata.

3*. Number of references to course materials and other sources
external to the course: includes explicit references by stu-
dents to course materials within and outside the course e.g.,

slide 4, from wikipedia, lecture video 7.

4*. Affirmations by fellow students; Count of agreements made
by fellow students in response to a post. Mostly, first posts
in a thread receive affirmations.

5. Thread properties (Number of posts, comments, and both
posts / comments, Average number of comments per post):
expresses a thread’s length and structural properties in terms
of number of posts and comments posted.

6. Number of sentences in the thread: This feature intends to
capture long focussed discussions that may be intervened
more often than the rest.

7*. Number of non-lexical references to course materials: (num-
ber of URLs, references to time points in lecture videos).
This feature is similar to course material references but in-
cludes only non-lexical references (Item #1) such as URLs
time points in lecture videos.

Importantly, as part of the author information, Coursera also marks
instructor-intervened posts / comments. This supplies us with auto-
matically labeled gold standard data for both training and evaluat-
ing our supervised classifier. We use threads with instructor posts /
comments as positive instances (intervened threads). However, we
note that the class imbalance is significant: as the instructor-student
ratio is very low, typical MOOC forums have fewer positives (in-
terventions) than negative ones. To counter skewness, we weigh

3Some courses had forums for projects, labs, peer assessment, dis-
cussion assignments. We omit from the collection these and other
miscellaneous forums, such as those for general discussion, study
groups and technical issues.

4Previous work such as [5] utilize this as they have access to time-
stamped versions of these statistics, since they use privately-held
data supplied by Coursera for MOOCs held at their university.
4. EVALUATION

We performed detailed experimentation over the smaller D14 dataset to validate performance, before scaling to the D61 dataset. We describe these set of experiments in turn. As our task is binary classification, we adopted L1-regularized logistic regression as our supervised classifier in all of our experimentation.

We first investigated each of the 14 courses in D14 as 14 separate experiments. We randomly used 80% of the course’s threads for training and validation (to determine the class weight parameter, \( W \)), and use the remaining 20% for testing. Our experimental design for this first part closely follows the previous work [5] for direct comparison with their work. We summarise these results in Table 2, in the columns marked “(II) Individual”, averaging performance over ten-fold cross validation for each course.

The results show a wide range in prediction performance. This casts doubt on the portability of the previously published work [5]. They report a baseline performance of \( F_1 \approx 25 \) on both their courses each having an intervention ratio \( \approx 0.13 \). In contrast, our results show the instability of the prediction task, even when using individualized trained models. Nevertheless, on average our set of features performs better on \( F_1 \) by at least 10.15%.

We observe the true intervention ratio correlates to performance, when comparing Columns I.2 and II.3 (\( \rho = 0.93 \)). We also see that intervention ratio varies widely in our D14 dataset (Figure 4). This happens to also hold for the larger D61 dataset. In some courses, instructors intervene often (76% for medicalneuro-002) and in some other courses, there is no intervention at all (0% for biostats-005). To see whether the variability can be mitigated by including more data, we next perform a leave-one-course-out cross validation over the 14 courses, shown in “Columns (III) LOO-course C.V.”. I.e., we train a model using 13 courses’ data and apply the trained model to the remaining unseen test course. While not strictly comparable with (II), we feel this setting is more appropriate, as it: allows training to scale; is closer to the real scenario discussed in Section 6, Item 4.

Separately, we studied the effectiveness of our proposed set of features over the D14 dataset. Table 3 reports performance averaged over all 14 courses weighted by its proportion in the corpus. In the top half of the table, we build Systems 1–7 by cumulatively adding in features from the proposed list from Section 3.2. Although the overall result in Row 7 performs \( \approx 5\% \) better than the unigram baseline, we see that the classifier worsens when the count of course references are used as a feature (Row 2). Other rows all show an additive improvement in \( F_1 \), especially the forum type and non-lexical reference features, which boost recall significantly.

The performance drop when adding in the number of course references prompts us to investigate whether removing some features from the full set would increase prediction quality. In the bottom half of Table 3, we ablate a single feature from the full set.

Results show that removing forum type, number of course references and thread length in a thread all can improve performance. Since the different rows of Table 3 are tested with weights \( W \) learnt from its own training set the changes in performance observed are due to the features and the learnt weight. When we tested the same sequence with an arbitrary constant weight we observed all features but Course_Ref improved performance although not every improvement was significant.

Using the best performing feature set as determined on the D14 experiments, we scaled our work to the larger D61 dataset. Since a leave-one-out validation of all 61 courses is time consuming we only test on the each of the 14 courses in D14 dataset while training on the remaining 60 courses from D61. We report a weighted averaged \( F_1 = 50.96 \) (\( P = 42.80; R = 76.29 \)) which is less than row 9 of Table 3. We infer that scaling the dataset by itself doesn’t improve performance since \( W \) learnt from the larger training data no longer counters the class imbalance leaving the testset with a much different class distribution than the training set.

4.1 Upper bound

The prediction results show that forum type and some of our newly-proposed features lead to significant improvements. However, we suspect the intervention decision is not entirely objective; the choice to intervene may be subjective. In particular, our work is based on major class (generally positive) instances higher than minority class (generally negative) instances. These weights are important parameters of the model, and are learned by optimizing for maximum \( F_1 \) over the training / validation set.

Table 2.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Precision</th>
<th>Recall</th>
<th>( F_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Unigrams</td>
<td>41.98</td>
<td>61.39</td>
<td>45.58</td>
</tr>
<tr>
<td>2. (1) + Forum Type</td>
<td>41.36</td>
<td>69.13</td>
<td>48.01</td>
</tr>
<tr>
<td>3. (2) + Course_Ref</td>
<td>41.09</td>
<td>66.57</td>
<td>47.22</td>
</tr>
<tr>
<td>4. (3) + Affirmation</td>
<td>41.20</td>
<td>68.94</td>
<td>47.68</td>
</tr>
<tr>
<td>5. (4) + T Properties</td>
<td>42.99</td>
<td>70.54</td>
<td>48.86</td>
</tr>
<tr>
<td>6. (5) + Num Sents</td>
<td>43.08</td>
<td>69.88</td>
<td>49.77</td>
</tr>
<tr>
<td>7. (6) + Non-Lex Ref</td>
<td>42.37</td>
<td>74.11</td>
<td>50.56</td>
</tr>
<tr>
<td>8. (7) – Forum Type</td>
<td>41.33</td>
<td>83.35</td>
<td>51.16</td>
</tr>
<tr>
<td>9. (7) – Course Ref</td>
<td>45.96</td>
<td>79.12</td>
<td>54.79</td>
</tr>
<tr>
<td>10. (7) – Affirmation</td>
<td>42.59</td>
<td>71.76</td>
<td>50.34</td>
</tr>
<tr>
<td>11. (7) – T Properties</td>
<td>40.62</td>
<td>84.80</td>
<td>51.35</td>
</tr>
<tr>
<td>12. (7) – Num Sents</td>
<td>42.37</td>
<td>73.05</td>
<td>49.32</td>
</tr>
<tr>
<td>13. (7) – Non-Lex Ref</td>
<td>43.08</td>
<td>69.88</td>
<td>49.77</td>
</tr>
</tbody>
</table>

Table 3: Feature study. The top half shows performance as additional features are added to the classifier. Ablation tests where a single feature is removed from the full set (Row 7) are shown on the bottom half. Performance given as weighted macro-average over 14 courses from a leave-one-out cross course validation over D14.
the premise that correct intervention follows the historical pattern of intervention (where instructors already intervene), and may not be where general pedagogy would recommend prediction. We recognize this as a limitation of our work.

To attempt to quantify this problem, we assess whether peer instructors with general teaching background could replicate the original intervention patterns. Three human instructors\(^7\) annotated 13 threads from the musicproduction-006 course. We chose this course to avoid bias due to background knowledge, as none of the annotators had any experience in music production. This course also had near zero interventions; none of the 13 threads in the sample were originally intervened by the instruction staff of the course.

They annotated 6 exam threads and 7 lecture threads. We found that among exam threads annotators agreed on 5 out of 6 cases. Among lecture threads at least two of three annotators always agreed. On 4 out of 7 cases, all three agreed. The apparently high agreement could be because all annotators chose to intervene only on a few threads. This corresponds to a averaged interannotator agreement of \(k = 0.53\). The annotators remarked that it was difficult to make judgements, that intervention in certain cases may be arbitrary, especially when expert knowledge would be needed to judge whether factual statements made by students is incorrect (thus requiring instructor intervention to clarify). As a consequence, agreement on exam threads that had questions on exam logistics had more agreement at \(k = 0.73\).

While only indicative, this reveals the subjectiveness of intervention. Replicating the ground truth intervention history may not be feasible – satisfactory performance for the task may come closer to the interannotator agreement levels; i.e., \(k = 0.53\) corresponding to an \(F_1\) of 53%. We believe this further validates the significance of the prediction improvement, as the upper bound for deciding intervention is unlikely to be 100%.

\(^7\)The last three authors, not involved in the experimentation: two professors and a senior pedagogy researcher.

5. DISCUSSION

From handling the threads and observing discussion forum interactions across courses, several issues arise that merit a more detailed discussion. We discuss each in turn, identifying possible actions that may mollify or address these concerns. Specifically:

1. The number of threads per course varies significantly.
2. Intervention decisions may be subjective.
4. Previous experimental results are not replicable.

**Issue 1: Variation in the number of threads.** We observed significant variation in the number of threads in different courses, ranging from tens to thousands. Figure 4 shows thread distribution over the D14 dataset for the errata, homework, lectures and exam forums; a similar distribution held for the larger D61 dataset. These distributions are similar to those reported earlier in the large cross-course study of [18]. The difference in number of threads across courses is due to a multitude of factors. These include number of students participating, course structure, assignment of additional credits to participating students, course difficulty, errors in course logistics and materials, etc.

When performing research that cuts across individual MOOCs, this issue becomes important. As we saw, using simple averaging on a per-course basis equates to a macro-averaging: putting each course on par in importance. However, when the decision unit is at the thread (as in our task), it makes more sense to treat individual threads at parity. In such cases, normalization at the thread level (analogous to micro-averaging) may be considered. Such thread-level normalization can affect how we weight information from each course when training in aggregate over data from multiple courses: courses with many threads should carry more weight in both training and evaluation.

**Issue 2: Intervention decisions may be subjective.** Instructor policy with respect to intervention can markedly differ. Instructors may only intervene in urgent cases, preferring students to do peer learning and discovery. Others may want to intervene often,

<table>
<thead>
<tr>
<th>Course</th>
<th>(I) Demographics</th>
<th>(II) Individual</th>
<th>(III) LOO-course C.V.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. # of Threads</td>
<td>2. I. Ratio</td>
<td>1. Prec.</td>
</tr>
<tr>
<td>ml-005</td>
<td>2058</td>
<td>0.45</td>
<td>51.08</td>
</tr>
<tr>
<td>rprog-003</td>
<td>1123</td>
<td>0.32</td>
<td>50.77</td>
</tr>
<tr>
<td>calc1-003</td>
<td>965</td>
<td>0.60</td>
<td>60.98</td>
</tr>
<tr>
<td>smac-001</td>
<td>632</td>
<td>0.17</td>
<td>21.05</td>
</tr>
<tr>
<td>compilers-004</td>
<td>624</td>
<td>0.02</td>
<td>8.33</td>
</tr>
<tr>
<td>maththink-004</td>
<td>512</td>
<td>0.49</td>
<td>46.59</td>
</tr>
<tr>
<td>medicalneuro-002</td>
<td>323</td>
<td>0.76</td>
<td>100.00</td>
</tr>
<tr>
<td>bioelectricity-002</td>
<td>266</td>
<td>0.76</td>
<td>75.00</td>
</tr>
<tr>
<td>bioinfomethods1-001</td>
<td>235</td>
<td>0.55</td>
<td>56.00</td>
</tr>
<tr>
<td>musicproduction-006</td>
<td>232</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>comparch-002</td>
<td>132</td>
<td>0.46</td>
<td>47.62</td>
</tr>
<tr>
<td>casebasedbiostat-002</td>
<td>126</td>
<td>0.20</td>
<td>13.33</td>
</tr>
<tr>
<td>gametheory2-001</td>
<td>125</td>
<td>0.19</td>
<td>28.57</td>
</tr>
<tr>
<td>biostats-005</td>
<td>55</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td>529</td>
<td>0.36</td>
<td>39.95</td>
</tr>
<tr>
<td>Weighted Macro Avg</td>
<td>NA</td>
<td>0.40</td>
<td>45.44</td>
</tr>
</tbody>
</table>

Table 2: Individual course results for each course in the D14 dataset. Weights \(W\) weigh each +ve class instance \(w\) times as much as a –ve class instance. Performance varies with large variations in Intervention ratio (I-ratio) and # of threads.
to interact with the students more and to offer a higher level of
guidance. Which policy instructors adopt varies, as best practices
for both standard classroom and MOOC teaching have shown both
advantages and disadvantages for [12, 11].

Instructors can also manifest in different roles. In Coursera, posts
and comments marked as instructor intervened can come from ac-
tual instructor intervention as well as participation by helpers, such
as community teaching assistants (CTAs). We observe courses with
CTAs where CTAs have a higher intervention rate. We hypothesize
that such factors decreases agreements.

This plays out in our datasets. We observe that intervention is not
always proportional to the number of threads in the course. Some
courses such as compilers-004 (see Figure 4) has relatively fewer
number of threads than other large courses. Yet its intervention rate
is noticeably low. This suggests that other factors inform the
intervention decision. Handling this phenomenon in cross-course
studies requires an additional form of normalization.

To normalize for these different policies we can upweight (by over-
sampling) threads that were intervened in courses with fewer in-
terventions. We can continue to randomly oversample a course’s
intervened threads until its intervention density reaches the dataset
average. Note this normalization assumes that the few threads in-
tervened in course with relatively low intervention density are more
important; that the threads intervened for a similar high interven-
tion density course would be a proper superset.

Even when a policy is set, intervention decisions may be subjective
and non-replicable. Even with our cursory annotation of a course
to determine an upper bound for intervention shows the potentially
large variation in specific intervention decisions. We believe that
automated systems can only approach human performance when
such decisions can be subjective. As such, the upper bound for per-
formance (cf Section 4.1) should not be construed as the single
gold standard; rather, prediction performance should be calibrated
to human performance levels.

Issue 3: Simple baselines outperform learned models. We also
compared our work with a simple baseline that predicts all threads
as needing instructor intervention. This baseline does no work –
achieving 100% recall and minimal precision – but is very compet-
titive, outperforming our learned models for courses with high levels
of intervention (see Table 4). Diving deeper into the cause, we at-
tribute this difference to the subjective nature of interventions and
other extraneous reasons (bandwidth concerns) resulting in high
false positive rates. That is, given two threads with similar set of
features, one may be intervened while the other is not (e.g., Fig-
ure 5). This makes the ground truth and the evaluation less reliable.
An alternative evaluation model might be to assign a confidence
score to a prediction and evaluate the overlap between the high con-
fidence predictions and the ground truth interventions.

Issue 4: Previous results are not replicable. From earlier work
[5], intervention prediction seemed to be straightforward task where
improvement can be ascribed to better feature engineering. How-
ever, as we have discovered in our datasets, the variability in in-
structor intervention in MOOCs is high, making the application of
such previously published work to other MOOCs difficult. This is
the perennial difficulty of replicating research findings. Find-
ings from studies over a small corpus with select courses from spe-
cific subject categories may not generalise. Published findings are
not verifiable due to restricted access to sensitive course data. The

<table>
<thead>
<tr>
<th>Course</th>
<th>Individual F1</th>
<th>F1 @100R</th>
<th>DI4 F1</th>
<th>F1 @100R</th>
</tr>
</thead>
<tbody>
<tr>
<td>ml-005</td>
<td>64.96</td>
<td>63.79</td>
<td>72.35</td>
<td>61.83</td>
</tr>
<tr>
<td>rprog-003</td>
<td>49.62</td>
<td>47.39</td>
<td>48.55</td>
<td>49.31</td>
</tr>
<tr>
<td>calc1-003</td>
<td>51.29</td>
<td>74.83</td>
<td>70.63</td>
<td>75.33</td>
</tr>
<tr>
<td>smce-001</td>
<td>25.00</td>
<td>34.07</td>
<td>34.15</td>
<td>29.28</td>
</tr>
<tr>
<td>compilers-004</td>
<td>14.28</td>
<td>3.28</td>
<td>4.82</td>
<td>4.75</td>
</tr>
<tr>
<td>mathblank-004</td>
<td>63.56</td>
<td>63.08</td>
<td>61.11</td>
<td>65.49</td>
</tr>
<tr>
<td>medicalneuro-002</td>
<td>75.36</td>
<td>86.66</td>
<td>78.06</td>
<td>85.67</td>
</tr>
<tr>
<td>bioelectricity-002</td>
<td>63.16</td>
<td>86.84</td>
<td>80.10</td>
<td>85.84</td>
</tr>
<tr>
<td>bioinfomethods1-001</td>
<td>58.33</td>
<td>67.65</td>
<td>69.40</td>
<td>71.07</td>
</tr>
<tr>
<td>musicproduction-006</td>
<td>0.00</td>
<td>4.35</td>
<td>1.09</td>
<td>1.72</td>
</tr>
<tr>
<td>comparch-002</td>
<td>64.57</td>
<td>55.56</td>
<td>60.49</td>
<td>63.21</td>
</tr>
<tr>
<td>casebasedbiostat-002</td>
<td>23.53</td>
<td>14.81</td>
<td>38.71</td>
<td>34.25</td>
</tr>
<tr>
<td>gametheory2-001</td>
<td>28.57</td>
<td>45.16</td>
<td>27.12</td>
<td>30.56</td>
</tr>
<tr>
<td>biostats-005</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td>41.59</td>
<td>46.43</td>
<td>45.18</td>
<td>47.09</td>
</tr>
<tr>
<td>Weighted Macro Avg</td>
<td>49.04</td>
<td>51.51</td>
<td>54.79</td>
<td>53.22</td>
</tr>
</tbody>
</table>

Table 4: Comparison of F1 in Table 2 with those of a naïve
baseline that classifies every instance as +ve – resulting in 100% recall.

The main challenge is to provision secured researcher access to the
experimental data. Even in cases where researchers have access
to larger datasets, such prior research [1, 5, 14, 15, 16, 22] have
reported findings on each course separately (cf Table 2 “Individual”),
skipping away from compiling them into a single dataset
in their study. Bridging this gap requires cooperation among in-
terested parties. The shared task model is one possibility: indeed,
recently Rose et al. [17] organised a shared task organised to pre-
dict MOOC dropouts over each week of a MOOC. To effectively
make MOOC research replicable, data must be shared to allow oth-
ers to follow and build on published experimentation. Similar to
other communities in machine learning and computational linguistics,
the community of MOOC researchers can act to legislate data
sharing initiatives, allowing suitably anonymized MOOC data to be
shared among partner institutes.

We call for the community to seize this opportunity to make re-
search on learning at scale more recognizable and replicable. We
have gained the endorsement of Coursera to launch a data-sharing
initiative with other Coursera-partnered universities. While we rec-
ognize the difficulties of sharing data from the privacy and institu-
tional review board perspectives, we believe that impactful research
will require application to a large and wide variety of courses, and
that restricting access to researchers will alleviate privacy concerns.

6. A FRAMEWORK FOR INTERVENTION RESEARCH

We have started on the path of instructor intervention prediction,
using the task formalism posed by previous work by Chaturvedi
et al. [5]: the binary prediction of whether a forum discussion
thread should be intervened, given its lexical contents and meta-
data. While we have improved on this work and have encouraging
results, this binary prediction problem we have tackled is overly
constrained and does not address the real-world need for interven-
tion prediction. We outline a framework for working towards the
real-world needs of instructor intervention.
We thus propose a framework for investigation that iteratively relaxes our problem to take into account successively more realistic aspects of the intervention problem, with the hope of having a fieldable, scalable, real-time instructor intervention tool for use on MOOC instructors’ dashboard as an end result.

1. Thread Ranking. We posit that different types of student posts may exhibit different priorities for instructors. A recommendation for intervention should also depend on thread criticality. For example, threads reporting errors in the course material may likely be perceived as critical and hence should be treated as high-priority for intervention. Even with designated errata forums, errata are reported in other forums, sometimes due to the context – e.g., when a student watches a video of a lecture, it is natural for him to report an error concerning it in the lecture forum, as opposed to the proper place in the errata forum. Failure to address threads by priority could further increase the course’s dropout rate, a well-known problem inherent to MOOCs [6]. Thread ranking can help to address this problem to prioritize the threads in order of urgency, which the naïve, always classifying all instances as positive, baseline system cannot perform.

2. Re-intervention. Threads can be long and several related concerns can manifest within a single thread, either by policy or by serendipity. Predicting intervention at the thread level is insufficient to address this. A recommendation for intervention has to consider not only those threads that had been newly-created but also if older threads that had already been intervened require further intervention or reinvention. In other words, intervention decision needs to be made in the light of newly posted content to a thread. We can change the resolution of the intervention prediction problem to one at the post level, to capture re-interventions; i.e., when a new post within a thread requires further clarification or details from instructor staff.

3. Varying Teaching Roles. MOOCs require different instruction formats than the traditional course format. One evolution of the MOOC teaching format to adapt to the large scale is to recruit community teaching assistants (CTAs). Community TAs are volunteer TAs recruited by MOOC platforms including Coursera based on their good performance in the previous iteration of the same MOOC. CTAs, traditional Teaching Assistants and technical staff are all termed as “staff” within the Coursera system. Currently, Coursera only marks threads with a “staff replied” marking, which we use directly in our training supervision in this paper. At a post level, those posted by CTAs, instructor and technical staff are marked appropriately.

We hypothesize that that these various roles differ in the quantum of time and effort, and type of content that they provide in answering posts that they contribute on a forum. It will be important to consider the role of the user while recommending threads to intervene, as the single problem of intervention may lead to separate triaging problems for the n staff types or individual instructors that manage a MOOC.

4. Real-time. In the real world, a system needs to be predicting intervention in real-time; as new posts come into a course’s forum. With ranking, we can decide when to push notifications to the instruction staff, as well as those less urgent that can be viewed at leisure on the instructor’s MOOC dashboard.

With the timestamp metadata in the dataset, we have a transaction log. This allows us to easily simulate the state of a MOOC by “rewinding” the state of the MOOC at any time t, and make a prediction for a post or thread based on the current state.

This half-solves the problem. For real-world use, we also need to do online learning, by observing actual instructor intervention and adopting our system for the observed behavior. We feel this will be important to learn the instructor’s intervention preference, as we have observed the variability in intervention per course, per instructor.
In our work, we focus only on the instructor’s view, however this set of problems also has an important dual problem set: that of the student’s view. We believe that solving both problems will have certain synergies but will differ in important ways. For example, solving the student’s view will likely have a larger peer and social component than that for instructors, as MOOCs develop more social sensitivity.

7. CONCLUSION
We describe a system for predicting instructor intervention in MOOC forums. Drawing from data over many MOOC courses from a wide variety of coursework, we devise several novel features of forums that allow our system to outperform the state-of-the-art work on an average by a significant margin of 10.15%. In particular, we find that knowledge of where the thread originates from (the forum type – whether it appears in a lecture, homework, examination forum) alone informs the intervention decision by a large 2% margin.

While significant in its own right, our study also uncovers issues that we feel must be accounted for in future research. We have described a framework for future research on intervention, that will allow us to account for additional factors – such as temporal effects, differing instructor roles – that will result in a ranking of forum threads (or posts) to aid the instructor in best managing her time in answering questions on MOOC forums.

Crucially, we find the amount of instructor intervention is widely variable across different courses. This variability undermines the veracity of previous works and shows that what works on a small variable across different courses. This variability undermines the veracity of previous works and shows that what works on a small scale may not hold well in large, cross-MOOC studies. Our own results show that for many courses, simple baselines work better than supervised machine learned models when intervention ratios are high. To allow the replicability of research and to advance the field, we believe that MOOC-fielding institutions need to form a data consortium to make MOOC forum data available to researchers.

8. ACKNOWLEDGMENTS
The authors would like to thank Snigdha Chaturvedi and her co-authors for their help in answering detailed questions on the methodology of their work.

9. REFERENCES