QuanTyler : Apportioning Credit for Student Forum Participation

Ankita Bihani
Stanford University
Stanford, USA
ankitab@stanford.edu

Andreas Paepcke
Stanford University
Stanford, USA
paepcke@cs.stanford.edu

ABSTRACT

We develop a random forest classifier that helps assign academic credit for a student's class forum participation. The classification target are the four classes created by student rank quartiles. Course content experts provided ground truth by ranking a limited number of post pairs. We expand this labeled set via data augmentation. We compute the relative importance of the predictors, and compare performance in matching the human expert rankings. We reach an accuracy of 0.96 for this task. To test generality and scalability, we trained the classifier on the archive of the Economics Stack Exchange reputation data. We used this classifier to predict the quartile assignments by human judges of forum posts from a university Artificial Intelligence course. Our first attempt at transfer learning reaches an average AUC of 0.66 on the augmented test set.

Keywords

Online Discussion Forum, MOOCs, residential courses, random forest, credit computation, online learning, transfer learning, instructor support, collaborative learning, grading, crowdsourcing, forum assessment.

1. INTRODUCTION

Massively Open Online Courses (MOOCs) have in past years provided content to populations outside traditional venues of higher education. For these settings, online forum facilities that are built into the course delivery platforms, such as Coursera and Open edX are the primary means of communication among learning peers, and for interacting with instructors.

Beyond the practical needs for coordinating logistics in geographically distributed settings, online discussion forums can serve pedagogical goals as well. Online asynchronous discussion forums provide the basis for collaborative learning, which enhances critical thinking [10]. Students answering each others' questions can be helpful for all parties [14].

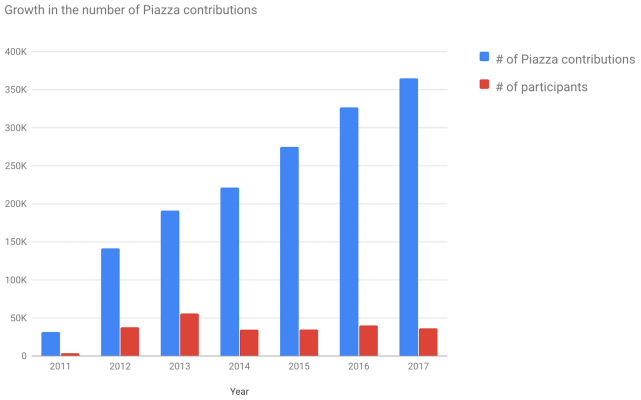

Figure 1: Number of Piazza forum contributions and participants per year for courses at our University

This support function is particularly useful in Science and Engineering courses. But as discussion centric humanities courses embrace distance learning, discussions on online forums will likely gain even more prominence.

However, it is not only in the context of distance learning that forum facilities have found uses. Even when in person class time is available, many residential college courses have adopted the tool. The need for students to ask questions, voice concerns, or to point out errors in course material are as salient in residential settings as they are in less traditional situations, such as distance learning [4]. Figure 1 shows the rapid growth in the volume of contributions per year to Piazza, just one of the several available online forum tools in a large private university. Despite the fact that the total number of Piazza participants were roughly the same from 2012 to 2017 (with a slight peak in 2013), the total volume of contributions increased monotonically. It is possible that this rapid increase in the volume of contributions per year on Piazza stems from the increasing popularity and growing adoption of the Piazza forum among students and instructors for collaborative discussions.

Given the importance of collaborative discussion in the learning process at both the theoretical and empirical level, instructors in at least some universities are assigning between 1% and 25% of their course grading component to online forum contribution. Two primary challenges arise when apportioning course credit to reward students' forum contri-

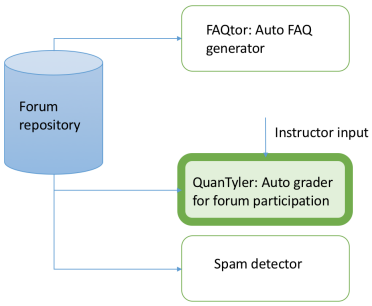

Figure 2: Block diagram of our proposed framework around forum facilities.

bution. First, students can attempt to game the system. On surveying some instructors, we learnt about instances of students copying a peer's forum posts, adding spaces or other innocuous characters to fool automated contribution counters. Thus, the system needs to be able to flag such instances.

A second, more complicated problem is that of apportioning fair credit to the students at the end of the quarter. Forum contributions take many forms. Asking an insightful or intriguing question can contribute as much to the course as providing answers. Taking the time to view other students' contributions is a contribution as well. However, for courses with hundreds of students, manual assessment of every forum post by each student in order to assign a forum participation score is not feasible. On surveying some instructors, we learnt that they instead develop ad–hoc formulae over the participation statistics provided by the forums, hoping to capture the right signals. This practice can not only lead to non-uniform grading (based on diverging intuitions) across courses, but also fail in rewarding students with a fair forum participation credit commensurate with their effort.

In addition to the above two challenges, there is untapped potential from today's use of forum facilities. As courses are offered repeatedly over the years, a treasure of course knowledge accumulates in forum archives. The detection of high value forum contributions can inform content selection from such archives.

In an effort to address these problems we are developing a coherent system for boosting the value of online discussion forums. Figure 2 shows a block diagram of our proposed system. In this paper, we focus on *QuanTyler*, the module responsible for helping with automatic forum credit apportioning. This component is highlighted in the figure. We plan to make the operation of *QuanTyler* customizable by instructors. For instance, instructors will be able to decide the granularity of partitioning the class into their quantiles of choice.

We begin with describing how we used human judgments to establish ground truth of what a 'good' and credit-worthy forum contribution looks like. This ground truth is used for measuring success, and for training the models. At the heart of our contribution are three experiments whose outcomes are required to inform the development of the *QuanTyler* module. These experiments are outlined below.

In the first experiment, we explore how students can be classified into quantiles based on their forum contributions, such that the implied ranking matches the ground truth. We show the hyperparameters needed to make a Random Forest classifier work well in support of the post evaluation task. We reached a high AUC measure in this task.

However, obtaining human judgments is expensive. At the same time, this requirement for human judgment would limit the ability to create classifiers for many courses. To break out of this confinement, we examine how a much larger source of labels for a forum-like enterprise might be used for training, and to test generalizability.

To this purpose we used *Stack Exchange*, [2] which is an online Q & A platform with millions of users. Stack Exchange is partitioned into sites for varying disciplines. We chose the Economics archive [1], and used it as a source for attempting transfer learning. In our second experiment we trained a random forest model on Stack Exchange reputation data, and tried predicting the quality ratings of human expert-rated forum posts in an Artificial Intelligence (AI) class. While not as good a classifier as the one trained on the forum data itself, this first attempt at transfer learning reached an $AUC = 0.66$, which we hope to improve further going forward. However, the data from Stack Exchange cannot be used in its raw form to build a classifier, and we will cover the required processing in Section 8.

In our third experiment, we demonstrated that (at least one of) the ad–hoc formulas currently deployed at our university diverges significantly from human experts' judgment.

2. RELATED WORK

Online discussion forums empower students and instructors to engage one another in ways that promote critical thinking, collaborative problem solving, and knowledge construction [20, 17]. Research has shown that linking some form of assessment to forum participation is an important element in promoting and enhancing online interactivity [16, 28].

Quantitative methods for content analysis are most widely used in assessing effective forum participation. [7] presents an overview of 15 different content analysis instruments used in computer supported collaborative learning (CSCL) studies.

The model proposed by [12] is a common starting point in many CSCL studies. In [12], the author presented a framework and analytical model to evaluate computer-mediated communication (CMC). The analytical model was developed to highlight five key dimensions of the learning process exteriorized in messages: *participation, interaction, social, cognitive and metacognitive dimensions.* Although this model provides an initial framework for coding CMC discussions, it lacks detailed criteria for systematic and robust classification of electronic discourse [13].

Many researchers have strongly endorsed Social Network

Analysis as a key technique in assessing the effectiveness of forum interactions [6, 29, 8]. Social Network Analysis is a research methodology that seeks to identify underlying patterns of social relations based on the way actors are connected with each other [25, 22].

In [18], the authors discuss a conceptual framework for assessing quality in online discussion forums. Drawing on previous work [12, 19, 9], the authors propose three broad categories of criteria for assessing forum participation: *content*, which demonstrates the type of skill shown by the learners, *interaction quality*, which looks at the way learners interact with each other in a constructive manner, and *objective measures*, which highlight the frequency or participation. These three broad criteria are further divided, resulting in a total of 11 criteria. In order to support educators, the framework outlines a further sub classification, clearly indicating what may be a poor, satisfactory, good or excellent performance against each criterion. The primary limitation of this study is that manual assessment by instructors is not feasible in courses with hundreds of students.

In [24], the authors adopt a content analysis approach and develop a coding scheme to analyze students' discussion behaviors in order to categorize them as active, constructive or interactive. However, the authors do not discuss how to apportion forum participation credit based on the behaviors depicted. One of their findings shows that higher quantity of participation in the MOOC discussion forums is associated with higher learning gains. In coherence with this finding, we also include participation count as one of our potential predictors.

To the best of our knowledge, the most closely related work to our paper are [21] and [23].

In [21], the authors present the use of Social Network Analysis (SNA) to examine the structure and composition of ties in a given network, and provide insights into its structural characteristics. In particular, the authors rely on two types of networks: interaction network between students in a course, and the network of terms used in their interactions. The dynamic visualization of interaction between participants and the groups or communities formed can help the instructors rank students based on their centrality in the students' interaction network. Visualizing the network of terms used in an online discussion forum can be used to compare the interest of different students and their relative engagement.

In [23], the author proposes the use of the following metrics to assess forum participation: initiative, effectiveness–depth, effectiveness–breadth, value, timeliness, participation, scholarship, style, and instructor points. Our system explicitly or implicitly covers most of these measures and augments them further by adding the crucial element of social network analysis to assess forum participation.

In contrast with both the aforementioned contributions, each of which focuses on specific aspects for assessing forum participation, our approach for assessing a student's contribution uses a combination of quality measures, quantitative measures, engagement level measures and also measures from social network analysis. The intent is to provide a holistic view of each student's contribution. We develop a system that the instructors can customize and easily use for apportioning forum participation credit.

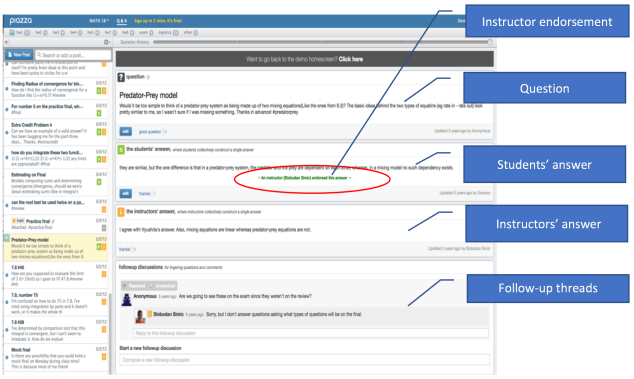

Figure 3: Sample annotated screenshot of the Piazza forum facility.

3. CURRENT PRACTICE

Many universities use the Piazza forum facility [27] for asynchronous online discussions. In order to provide context for the experiments below, we provide a brief overview of this tool.

Piazza is a Q&A web service for online discussions, where users can ask questions, answer questions, and post notes. The user interface contains a dynamic list of posts, which are question titles followed by a snippet of lines from the post. For every question, there is a placeholder for the instructor's answer, which can only be edited by instructors. There is also a students' answer section where students *collaborate* to construct a single answer. Students can *upvote* each others' questions or answers. Instructors can also *endorse* good questions and answers, which are then highlighted as instructor endorsed. There is also a discussion segment for follow-up threads. Figure 3 shows a snapshot of the Piazza discussion forum.

On surveying several instructors who consider Piazza forum participation in their grading scheme, we found that most rely on the basic quantitative statistics that the forum machinery readily offers. The following were some of the grading schemes that are currently used by instructors at our university for awarding forum participation grades:

Scheme 1: In this scheme, scores of each student were calculated using the following formula:

$$\begin{aligned}\texttt{Score} = 1 * \texttt{(no\_questions\_asked)} + \\ 4 * \texttt{(no\_questions\_answered)} + \\ 0.5 * \texttt{(other\_contributions)}\end{aligned} \quad (1)$$

Scheme 2: In this scheme, scores of each student were calculated using the following formula:

$$\begin{aligned}\texttt{Score} = 3 * \texttt{(no\_questions\_answered)} + \\ 1 * \texttt{(no\_followups)}\end{aligned} \quad (2)$$

Anyone above the 90^{th} percentile received full credit, and all the other students received a score of 0.
Scheme 3: Award full credit if at least one forum contribution was made, and the student was online on the forum for at least x number of days, or viewed at least y posts. Here x and y were set by the instructor using intuition.
Scheme 4: Award full credit to a student if they made at least one contribution to the forum.

All the above methods rely solely on the basic statistics directly provided by Piazza [27]. The concern, however, is whether these methods accurately and meaningfully award credit to the deserving students. Following are two major limitations of using the current grading schemes:

- *Lack of quality measures :* All the 4 grading schemes described above overlook the quality of contributions. This exclusion negatively impacts the grades of the students who post few, but very high quality contributions. More importantly, relying solely on the quantity of the contributions encourages posts that do not constitute meaningful forum participation. This behavior, in turn, can cause the forum's quality to devolve.
- *Reward not proportionate to effort :* Most of these schemes fail to award credit proportionate to the amount of effort and time the student invested. For instance, using the third or fourth scheme means that two students with vastly varying quantity and quality of contributions would be awarded the same score. Concretely, let us consider two students A and B. Student A made only one forum contribution during the course by posting a "+1" to another student's question. However, student B regularly made meaningful forum contributions throughout the quarter. Using grading scheme 4, both would receive equal credit. This lack of fairness can deter students from engaging in meaningful forum contributions.

Despite the above limitations, instructors have no choice but to rely on grading schemes like the ones discussed above. The large volumes of forum posts that accumulate by the end of the term make it impossible for the course staff to manually go through them to apportion credit. However, even if hypothetically, one were to have the course staff manually go through each of the contributions, there is a significant amount of subjectivity in assessing forum contributions. Having TAs manually grade contributions would lead to a lack of grading uniformity. A trivial contribution according to one TA, might be a significant contribution to another. Thus, there is a need for an automated way to assess the forum participation of students using a holistic grading scheme. Automation can lead to a standardized approach across the entire class.

The next sections discuss how we developed a system to assess forum participation by each student at scale. We go beyond the ready at hand statistics that are provided by the forum, and additionally incorporate measures that provide insight into the dynamics of students interacting in the forum. These dynamics manifest in the social networks that are created by the online interactions. We briefly review candidate predictors in the next section.

4. POTENTIAL PREDICTORS

The measurements of predictors arise from the data sets generated by forum facilities during the length of an academic term. Each offering of a course generates a separate data set, such as the one we used from the AI course.

Quantitative measures: These measures reward based on the volume of contributions made by an individual. As discussed in [24], higher forum participation count translates to higher learning gains, hence we include quantitative features in our list of potential predictors. These four predictors are: *number of questions asked*, *number of questions answered*, *total number of contributions*, and *average post length* by a student.

Engagement level measures: In order to reward the students who started important or intriguing threads, which in turn engaged many students, the *average number of collaborators* in the threads started by the student was added as a predictor. A second predictor, *average number of views* received by a student's questions was added for similar reasons. Given that not everyone in the class might be comfortable actively posting on the forum, we use some metrics to reward the passive engagement of the students. Some of the students are great listeners; they view or follow most of the posts, and are regularly online on the forum, which translates to passive forum participation. The two predictors we used to apportion credit for passive collaboration on the forum are: *total number of days a student was logged into the forum*, and the number of *posts viewed* by the student.

Quality measures: These measures are used to reward the students based on the quality of their contributions. These include upvotes and endorsement counts available in forum datasets. Students can express appreciation for a post by adding an upvote to the contribution. Instructors can explicitly endorse answers provided by students, marking those answers as definitive. Upvotes and endorsements articulate human judgments, and can be thought of as crowdsourcing post quality assessment.

Another strength of quality measures is their robustness to student cheating by flooding the forum with meaningless threads to increase their contribution count. Our two quality predictors are: *number of endorsed answers by the student*, and *total number of endorsements*, including upvotes on the questions, answers and instructors' endorsements.

Social Network Analysis: As discussed in the Related work section, Social Network Analysis (SNA) provides insights to the student forum participation. A brief detour in the following section provides background for the measures we used for SNA.

In order to include the SNA component in our credit apportioning system, the following networks were extracted from the class forum dataset. In the definitions below, nodes represent students and instructors. Typed edges represent interactions that are possible in the forum. Link weights encode the number of such interactions between the link's nodes.

Upvotes network: An upvotes network is extracted, where an

edge from student A to student B indicates that A upvoted B's content at least once, and the weight of the edge encodes the number of times A upvoted B's content.

Endorsement network: An endorsement network is extracted, where an edge from instructor A to student B indicates that A endorsed B's content at least once, and the weight of the edge encodes the number of times A endorsed B's content.

Combined upvotes and endorsement network: This construct is a union of the above two networks. An edge from A to B indicates that A either upvoted and/or endorsed B's content at least once, and the weight of the edge encodes the sum of the upvotes and endorsements.

Interaction network: This graph models the interaction that happened on the forum over the duration of the course. In the interaction network, an edge from A to B indicates that B responded at least once to a question that A posted.

We use these networks to derive our final two predictors: *degree centrality* in the interaction network, and *page rank* in the combined upvotes and endorsement network.

We calculate the degree centrality for every node in the interaction network. Degree centrality measures the number of links incident upon a node. Higher degree centrality of a student implies that the student answered questions or resolved doubts for a large number of students. On a high level, degree centrality in the interaction network translates to the "helpfulness" and "resourcefulness" of the student. It also captures the breadth of the student's course knowledge.

Page rank in the combined upvotes and endorsement network was used in order to capture importance in both upvotes and endorsement information using a single metric. Page rank can additionally help uncover influential or important students in the network. Their reach extends beyond their immediate neighbors, and is therefore not captured by the earlier described upvote/endorsement measures. The higher the page rank in the combined network, the more "influential" the student.

5. GROUND TRUTH COLLECTION

In order to evaluate how effective each of the above predictors is in informing credit apportioning, we obtained human judgments by paying former students and teaching assistants of the AI or a related class to render judgments over a sample of posts. Given the high course enrollment of 700+, not all the posts could be evaluated. A survey instrument was used to collect the judgments, and participants were paid a $20 gift card. The number of posts sampled was limited by this cost, and time capacity of the 24 participants we could recruit.

Each item in the survey for the experts was a pair of two posts by different students. The experts were asked to indicate which of the two contributions was more helpful for the class as a whole. (See the precise instructions below). We chose this pairwise comparison method to economize on raters' time and attention, and because the derivation of full rankings from pairwise comparisons is well studied [11, 15].

Table 1: Kendall tau distance between rankings created by the 5 algorithms

	Algo1	Algo2	Algo3	Algo4	Algo5
Algo1	1	0.8538	0.2213	0.7243	0.2268
Algo2	0.8538	1	0.2132	0.6621	0.2050
Algo3	0.2213	0.2132	1	0.2306	0.3064
Algo4	0.7243	0.6621	0.2306	1	0.2741
Algo5	0.2268	0.2050	0.3064	0.2741	1

The task in preparing the survey was to find forum contribution pairs that would later help train an algorithm. The challenge was to select a set of posts that would cover a range of measures for all our candidate predictors, while being representative of the overall contributions. We describe here how this selection was accomplished.

Four algorithms use a weighted combination of the above explained candidate predictors.

- Alg 1: Using only quality measures and social network analysis measures.
- Alg 2: Using only quantitative measures and engagement level measures.
- Alg 3: Using all the measures with more emphasis placed on quantitative measures.
- Alg 4: Using all the measures with more emphasis placed on quality measures.

In addition, the current formula based grading scheme 1 that is used by some instructors at our university is included as a variant. Let us call this approach Alg 5:

$$\begin{aligned} \texttt{Score} = 1 * \texttt{(no\_questions\_asked)} + \\ 4 * \texttt{(no\_questions\_answered)} + \\ 0.5 * \texttt{(other\_contributions)} \end{aligned} \tag{3}$$

All the above five algorithms are then separately used to calculate each student's score. Table 1 shows the Kendall tau distance between the rankings created by each of the algorithms. Most of the values in the table are low, indicating low correlation between the rankings calculated by each of the 5 algorithms. This result is intuitive because all the 5 algorithms were designed by us to capture slightly different signals. As a next step, 10 new values are calculated, each of which are absolute values of ranking differences between one pair of rankings for the same student. Each algorithm pair is processed. Thus, we have Alg1vsAlg2, Alg1vsAlg3, Alg1vsAlg4, Alg1vsAlg5 and so on. For instance, if Student ID# 500 was ranked 30 by Alg 1, and 300 by Alg 3, then the Alg1vsAlg3 value for Student ID# 500 would be 270.

We then sort these ten rank differences in descending order. The top entry in the 10 columns gives us the corner cases, or students 'of interest'. To sample student pairs, we compare these students of interest with the students immediately above and immediately below in the ranking by both the algorithm rankings under consideration. We clarify the procedure with the following example:

Let us assume that student ID #10 had the maximum absolute difference between ranking through Alg 1 and Alg 3. Also, using Alg 1, student ID #400 is directly above student ID #10 and student ID #5 is directly below student ID #10 in the ranking. Finally, let us assume that using Alg 3, student ID #20 is directly above student ID #10 and student ID #557 is directly below student ID #10 in the ranking. Then posts by the student ID pairs of interest for which human judgment was solicited are: (10, 400), (10, 5), (10, 20), (10, 557).

In addition to 40 such pairs of interest, additional student pairs were randomly sampled. At most 4 question pairs and 4 answer pairs were sampled from all these selected student pairs and presented to the experts. A total of 89 question pairs and answer pairs were used. In order to avoid fatiguing the experts, the set was partitioned into two batches such that each question pair or answer pair was voted on by at least 12 experts. The set of judged samples thus served to inform boundary cases among available measures, rather than to include every type of post. For example, there was no attempt to cover all linguistic variations. The addition of randomly sampled posts served to reach beyond this focus.

The survey instructions were as follows:
Each of the following sections presents one pair of questions or answers that were posted to the course forum in the past. We ask that you to indicate for each pair, the contribution that might have been most helpful to the rest of the class.

One sample item from the survey is as follows:

Q1: I am very confused about alpha-beta pruning, as we do not have example code from lecture. When we say we prune certain leaf, what does it mean? Does it mean we do not store that choices?
Q2: To create our own label, must it been binary label {1,-1} or it can be multi-categories with labels of any number? Is the feature still word counts or can be anything?
Which of the above two questions contributes more to the class community?

Note that in all cases the experts who answered the surveys were different from the experts whose endorsements we counted when building the classifier.

In order to learn the experts' *intuition* about which of the predictors might be important in ranking students' forum contributions, the following related question was introduced in the ranking survey once at a random time, with a facility to drag the entries up and down to arrange the predictors in decreasing order of relevance:

Imagine you had the following statistics about forum contributions by all students at the end of the term. In your opinion, which statistics are important to evaluate the forum contributions of students to the class. Please drag the entries up and down to indicate their relative importance. The first entry would be the most important.

- *Number of questions asked by the student*
- *Number of questions answered by the student*
- *Total number of posts viewed by the student*
- *Total number of days the student was online on the forum*
- *Total number of endorsements received by the student*
- *Total number of Forum contributions by the student (including questions, answers, notes, follow-ups, etc.)*

Table 2: Experts' intuitions for relative ordering of indicator importance. Example: 57.1% of experts felt that the number of questions answered was the second-most important indicator.

Rank	Feature	%support
1	# of endorsements	60.7
2	# of questions answered	57.1
3	# of Forum contributions	46.4
4	# of questions asked	46.4
5	# of posts viewed	60.7
6	# of days online	64.2

Based on the majority vote for every rank, we arrive at a ranking order using the experts' intuition. This ranking was not used in any of the experiments below. The information just illustrates the 'gut' feeling by our raters. The results are summarized in Table 2. Rank 1 is the most important feature. The percentage of experts agreeing with each ranking is also included.

6. EXPERIMENT PREPARATION

Given the pairwise rankings of posts by the experts we needed to arrive at a ranking against which we could then train and test. We generated this ranking using the Copeland method [3]. The procedure counts the number of times a student's post was considered superior to the alternative post offered to an expert. The number of losses are then subtracted from these wins. Copeland ranking ties can be broken by a *second order Copeland* approach [5]. However, we found that forcing a complete order did not lead to good classification, because the ties are a reflection of true similarity.

We included at most 4 question pairs and 4 answer pairs from each sampled student in the survey. However, in most cases the sampled students had less than 4 questions / 4 answers. The final result was a list of 37 students for which we had rankings from twelve experts each. We collected this large number of rankings for each student because of the above mentioned subjectivity in evaluating posts. In addition to the rankings, we also had the measures for all our 12 candidate predictors for each of the 37 students .

Rather than attempting a regression, we formulated the problem as one of classification into four classes: the rank quartiles. This decision was based on the application of apportioning credit. A granularity of four suffices, given that forum participation is not the only source of credit for a course. Partitioning a 5% course credit into 700 values is not meaningful.

Given the sparsity of our human labeled set, we first augmented the labeled data as follows. We partitioned the ranked list of students into four roughly equal parts. Figure 4*a* shows the top two partitions using fictitious numbers

for clarity.

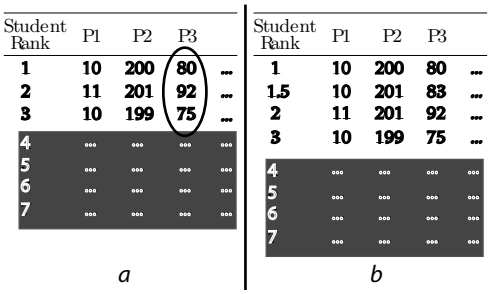

Figure 4: Augmentation occurs separately within each quantile. Each column holds the measures of one predictor P_n. The top two quantiles are shown. Part a: before augmenting the top quantile; Part b: after augmentation.

We then determined the range of values for each predictor within one quartile. Finally, we generated new rows within each quartile by randomly choosing values for each of its predictors from within the range of values that the predictor took on within that quartile.

The four quantiles could not be filled equally because of the ranking ties. Tied students should be in the same class, rather than being split across quantile boundaries. When such splitting occurred we moved all participants into one of the quantiles, such that the fewest moves were required. For example, if three of five students with rank seven were assigned to quartile two, and two were assigned to quartile three, all students ranked seven were moved to quartile two.

Finally, we set aside 30% of the resulting augmented set for testing. We call these sets *forumTrainAug* and *forumTestAug*. The corresponding putative responses are *forumTrainResp* and *forumTestResp*. Our first exploration was to see whether we could construct a classifier that would use predictor measures to assign each student to one of the quartiles.

7. EXPERIMENT 1: QUANTILE PREDICTION USING RANDOM FOREST

We started with a random forest (RF) of 10K trees in order to understand how many trees are required for this classification. Figure 5 shows the result of this investigation.

Table 3: Accuracy and Kappa by number of predictors per tree

mtry	Accuracy	Kappa
2	0.89	0.85
7	0.88	0.85
12	0.88	0.84

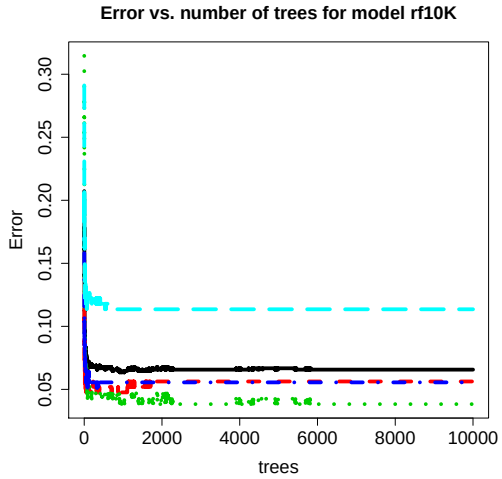

Figure 5: Classification errors by number of trees.

Each of the colored traces corresponds to one classifier. There are four traces, one corresponding to each quartile. The black line is the out-of-bag error. We see that after 6K trees the classification error no longer fluctuates. We settled on 8K trees to handle high data fluctuations. The second hyper parameter to tune, *mtry*, is the number of randomly chosen predictors that are used in each tree. The setting $mtry == 2$ was optimal, although this parameter is robust; see Table 3.

The resulting model *rf8K*, trained on *forumTrainAug* with 10-fold cross validation repeated 3 times has the confusion matrix shown in Table 4.

Table 4: Model RF8K predicting 308 augmented test set outcomes. Accuracy: 0.94

	RefQ1	RefQ2	RefQ3	Ref4	Class Error
PredQ1	76	0	2	0	0.03
PredQ2	1	77	0	14	0.16
PredQ3	0	0	75	0	0.00
PredQ4	1	0	1	62	0.02

Figure 6 shows the relative importance of our candidate predictors.

The chart shows the amount of decrease in accuracy that is contributed by each of the predictors. The top three predictors are the number of student answers that were endorsed by an instructor, the total number of endorsements, and the number of days the student was online on the forum. Note that these predictors differ somewhat from those intuited by

the experts, though there is some overlap.

Since there some of the predictors are partially covariant We experimented using three predictors only, but the degradation was noticeable. It is also advantageous to retain predictors that are less easy to spam than time online. For instance, the page rank predictor, while less important for the classification, is more difficult to defraud.

Using *rf8K* we predicted *forumTestResp*. Figure 7 shows ROC curves for each quartile predictor.

The prediction accuracy reaches 0.96. This result is encouraging in that it signals inroads towards apportioning fair forum participation credit even for very large courses.

However, the result does not speak to generality. The model was trained on a science forum data set, and its human labels were few. The classifier would not be useful if new labels needed to be created for each class. We therefore added a second experiment to demonstrate how the approach behaves when training occurs on data of an unrelated domain, and the resulting classifier is then used to predict forum participation ranking.

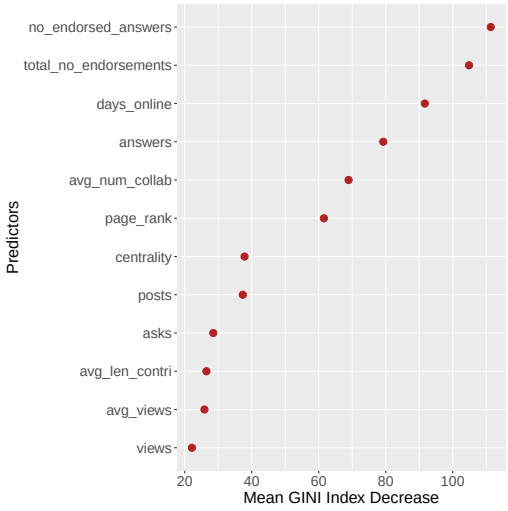

Figure 6: Mean decrease in GINI (node purity) when removing individual predictors. Ordered from most important at the top to least important.

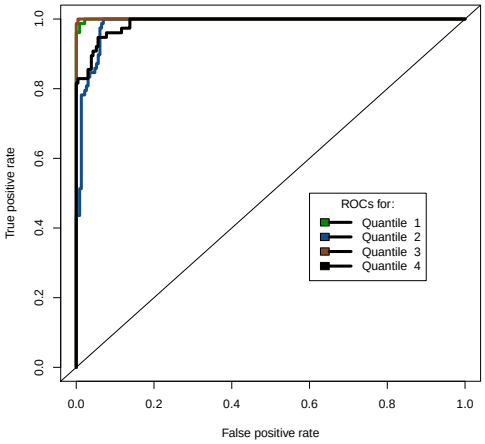

Figure 7: ROC curves for each quartile, predicted by 8000 random forest trees.

8. EXPERIMENT 2: STACK EXCHANGE TRANSFER LEARNING

Constructive activity on the Stack Exchange [2] forum earns users *reputation*, which can be used as a surrogate for forum participation credit. Among others, measures similar to the Piazza statistics we used in Experiment 1 are available from Stack Exchange, and we used those to predict reputation. However, only one of these measures is used by Stack Exchange for *their* computation of reputation; SE's algorithm instead takes six other variables into account.

We obtained the Stack Exchange (SE) records for the site dedicated to Economics [2].

We began with the data from about 5300 SE contributors. In a first step we followed the same procedure as in Experiment 1 to obtain optimal *mtry* and forest size values, which were 2 and 4K respectively. After scaling, centering, and partitioning into quartiles we set aside a 30% test set (*seTest*) from the training set (*seTrain*). The respective reputation responses are *seTrainResp*, and *seTestResp*.

Since the forum training set was not involved in the SE training, we used the larger *forumTrainResp* as test target for the SE-trained forest. Figure 8 shows the problematic resulting AUC ROC curves.

We addressed the lower triangle $Q3$ curve by reversing that classifier's orientation. This step is an appropriate measure, because the curve lies consistently below the diagonal, indicating a true polarity issue. However, AUC values were low, and further investigation uncovered the reason (Figure 9).

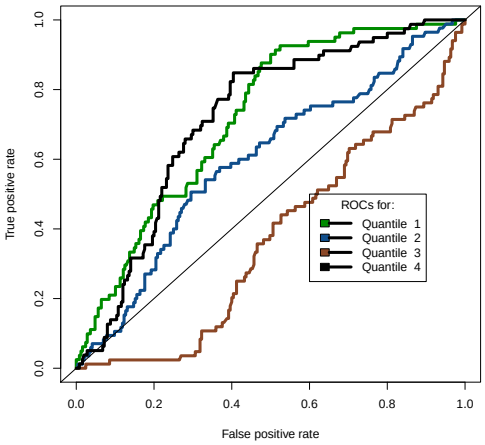

Figure 8: Initial AUC ROC from Stack Exchange-trained random forest predicting forum contribution quality.

Table 5: Confusion matrix for RF4K. OOB estimate of error rate: 27.81%

	Q1	Q2	Q3	Q4	Class error
Q1	502	25	83	313	0.46
Q2	3	859	1	34	0.04
Q3	149	6	716	29	0.20
Q4	126	230	9	539	0.40

Table 6: AUC Stack Exchange-trained model predicting forum post quality

	Q1	Q2	Q3	Q4	Mean
forumTrainResp	0.72	0.62	0.64	0.76	0.69
forumTestResp	0.78	0.64	0.45	0.77	0.66

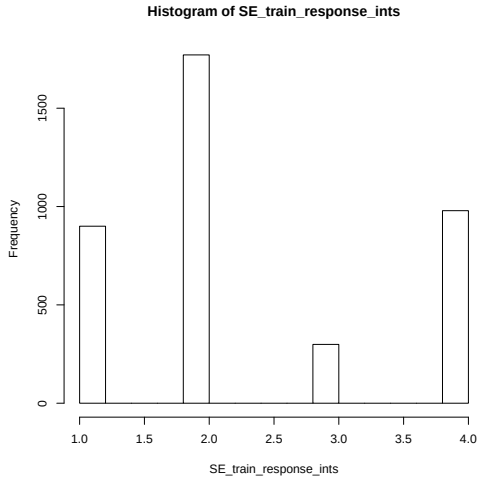

Figure 9: Class imbalance with raw Stack Exchange data

The Figure 9 shows that quartile 2 is over-represented, while quartile 3 suffers from a scarcity of examples. We balanced the training set by subsampling the quartile 2 examples to 1200, and augmented quartile 3 examples analogously to our process in Experiment 1.

The resulting 4K tree model, again trained with 10-fold cross validation repeated three times yielded a training accuracy of 0.72, and a kappa of 0.63. Table 5 shows the model's confusion matrix. When predicting *seTest* with this SE-trained classifier, a satisfactory mean AUC of 0.93 resulted, with classification behaviors shown in Figure 10.

Finally, with the SE model reasonably solid, we used this model to once again predict both *forumTrainResp* and *forumTestResp*. Table 6 shows results.

An important question remains: how do the ad hoc formulas devised by instructors perform? Are they sufficient? A final experiment tested the power of the informally designed Scheme 1 to approach the human expert judgments. Experiment 3 examines this question.

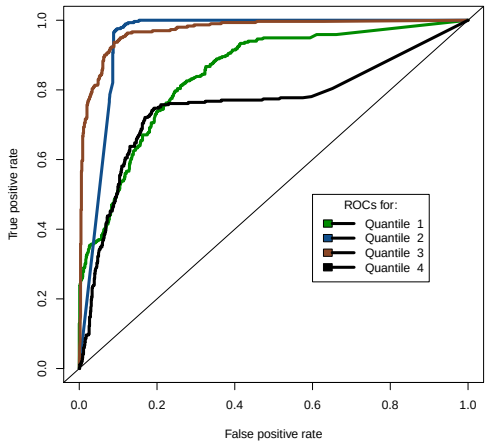

Figure 10: ROC for predicting Stack Exchange reputation from Stack Exchange-trained 4K random forest after attending to class imbalance.

9. EXPERIMENT 3: COMPARISON WITH CURRENT PRACTICE

We computed the quartile predictions induced by Equation 1, and compared them against $forumTestAug$ using Cohen's Kappa. This test returned a value of zero, evidence that the equation does not generate the same quartile assignments as the human experts. As a final check, we produced the categorical 1/0 quartiles for $forumTestAug$ from the $rf8K$ model, using 0.5 as the probability cutoff. The Cohen's Kappa between our model's prediction and the experts was 0.94.

10. DISCUSSION

The average AUC of 0.66 when using the Stack Exchange trained classifier on forum posts lags behind the classifier that is specialized on forum post evaluation. However, as a first step this result is encouraging. Forum assessment is gaining enough importance, and human judgments are expensive enough that training data from large, ready at hand, and similar enough facilities is extremely attractive for attempts in transfer learning.

Stack Exchange and other reputation incentivized systems have accumulated enough labeled samples that alternatives to random forests, such as neural nets, which require large amounts of training data might be feasible as approaches going forward.

11. CONCLUSION

Forum assessment is an active research area for good reason. A growing number of schools and companies are offering entire degree programs online, all of which require online communication among students and instructors. Demand for tools that help manage and assess forum activity is likely to rise as online education continues to capture market share.

Given that our attempt at transfer learning worked reasonably well, exploring the use of neural networks for automatic

forum participation grading is our next step. In addition, the work described here has not yet leveraged the content of the forum posts in assessing forum participation. In [26], the authors show that computational linguistic models can help in measuring learner motivation and cognitive engagement from the text of the forum posts. Hence, we plan to leverage Natural Language Processing techniques to analyze the content of the posts, and use those in apportioning forum participation credit. As explained in the introduction, this work is part of a larger effort that fills modules into a forum centered architecture. The frequently asked questions module and spam detection module will round out our efforts going forward.

12. REFERENCES

[1] Economics beta. World-Wide Web. Accessed Mar 6, 2018.

[2] Stack exchange data dump. World-Wide Web, 12 2017.

[3] A.H.Copeland. A âĂŸreasonableâĂŹ social welfare function. Notes from a seminar on applications of mathematics to the social sciences., 1951.

[4] M. A. Andresen. Asynchronous discussion forums: success factors, outcomes, assessments, and limitations. *Journal of Educational Technology & Society*, 12(1):249, 2009.

[5] W. contributors. Copeland's method — wikipedia, the free encyclopedia, 2016. [Online; accessed 7-March-2018].

[6] M. De Laat, V. Lally, L. Lipponen, and R.-J. Simons. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning*, 2(1):87–103, 2007.

[7] B. De Wever, T. Schellens, M. Valcke, and H. Van Keer. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & education*, 46(1):6–28, 2006.

[8] N. M. Dowell, O. Skrypnyk, S. Joksimovic, A. C. Graesser, S. Dawson, D. GaĹËĞevic, T. A. Hennis, P. de Vries, and V. Kovanovic. Modeling learners' social centrality and performance through language and discourse. *International Educational Data Mining Society*, 2015.

[9] D. R. Garrison, T. Anderson, and W. Archer. Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of distance education*, 15(1):7–23, 2001.

[10] A. A. Gokhale. Collaborative learning enhances critical thinking. *Journal of Technology Education*, 1995.

[11] R. Heckel, M. Simchowitz, K. Ramchandran, and M. J. Wainwright. Approximate ranking from pairwise comparisons. *CoRR*, abs/1801.01253, 2018.

[12] F. Henri. Computer conferencing and content analysis. In *Collaborative learning through computer conferencing*, pages 117–136. Springer, 1992.

[13] C. Howell-Richardson and H. Mellar. A methodology for the analysis of patterns of participation within computer mediated communication courses. *Instructional Science*, 24(1):47–69, 1996.

[14] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders. Superposter behavior in mooc forums. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 117–126, New York, NY, USA, 2014. ACM.

[15] K. G. Jamieson and R. D. Nowak. Active ranking using pairwise comparisons. *CoRR*, abs/1109.3701, 2011.

[16] D. Laurillard. *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge, 2013.

[17] R. M. Marra, J. L. Moore, and A. K. Klimczak. Content analysis of online discussion forums: A comparative analysis of protocols. *Educational Technology Research and Development*, 52(2):23, 2004.

[18] D. Nandi, S. Chang, and S. Balbo. A conceptual framework for assessing interaction quality in online discussion forums. *Same places, different spaces. Proceedings ascilite Auckland*, pages 7–23, 2009.

[19] D. R. Newman, B. Webb, and C. Cochrane. A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology*, 3(2):56–77, 1995.

[20] L. F. Pendry and J. Salvatore. Individual and social benefits of online discussion forums. *Computers in Human Behavior*, 50:211–220, 2015.

[21] R. Rabbany, S. ElAtia, M. Takaffoli, and O. R. Zaïane. Collaborative learning of students in online discussion forums: A social network analysis perspective. In *Educational data mining*, pages 441–466. Springer, 2014.

[22] J. Scott and P. J. Carrington. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[23] L. A. Tomei, editor. *Matthew Shaul: Information Communication Technologies for Enhanced Education and Learning: Advanced Applications and Developments: Advanced Applications and Developments*, chapter Assessing online discussion forum participation. IGI Global, 2008.

[24] X. Wang, D. Yang, M. Wen, K. Koedinger, and C. P. Rosé. Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. *International Educational Data Mining Society*, 2015.

[25] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[26] M. Wen, D. Yang, and C. P. Rosé. Linguistic reflections of student engagement in massive open online courses. In *ICWSM*, 2014.

[27] Wikipedia. *Piazza*, 2017.

[28] H.-T. Yeh. The use of instructor's feedback and grading in enhancing students' participation in asynchronous online discussion. In *Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on*, pages 837–839. IEEE, 2005.

[29] N. Yusof, A. A. Rahman, et al. Students' interactions in online asynchronous discussion forum: A social network analysis. In *Education Technology and Computer, 2009. ICETC'09. International Conference on*, pages 25–29. IEEE, 2009.