Generating Actionable Predictions regarding MOOC Learners' Engagement in Peer Reviews

Erkan Er [1], Eduardo Gómez-Sánchez [2], Miguel L. Bote-Lorenzo [3], Yannis Dimitriadis [4], Juan I. Asensio-Pérez [5]

GSIC/EMIC Research Group, Universidad de Valladolid, Valladolid, Spain

[1] **Corresponding author**: erkanererkaner@gmail.com| ETSI Telecomunicación, Paseo de Belén, 15 (Room 2L019), 47011 Valladolid, Spain. Tel. +34 983 42 3696 / 3698

[2] edugom@tel.uva.es | ETSI Telecomunicación, Paseo de Belén, 15 (Room 2L019), 47011 Valladolid, Spain. Tel. +34 983 42 3696 / 3698

[3] migbot@tel.uva.es | ETSI Telecomunicación, Paseo de Belén, 15 (Room 2L019), 47011 Valladolid, Spain. Tel. +34 983 42 3696 / 3698

[4] yannis@tel.uva.es | ETSI Telecomunicación, Paseo de Belén, 15 (Room 2L019), 47011 Valladolid, Spain. Tel. +34 983 42 3696 / 3698

[5] juaase@tel.uva.es | ETSI Telecomunicación, Paseo de Belén, 15 (Room 2L019), 47011 Valladolid, Spain. Tel. +34 983 42 3696 / 3698

This is an Accepted Manuscript of an article published by Taylor & Francis Group in Behaviour & Information Technology on 13/09/2019, available online:

https://www.tandfonline.com/doi/full/10.1080/0144929X.2019.1669222

Generating Actionable Predictions regarding MOOC Learners' Engagement in Peer Reviews

Abstract—Peer review is one approach to facilitate formative feedback exchange in MOOCs; however, it is often undermined by low participation. To support effective implementation of peer reviews in MOOCs, this research work proposes several predictive models to accurately classify learners according to their expected engagement levels in an upcoming peer-review activity, which offers various pedagogical utilities (e.g., improving peer reviews and collaborative learning activities). Two approaches were used for training the models: in situ learning (in which an engagement indicator available at the time of the predictions is used as a proxy label to train a model within the same course) and transfer across courses (in which a model is trained using labels obtained from past course data). These techniques allowed producing predictions that are actionable by the instructor while the course still continues, which is not possible with post-hoc approaches requiring the use of true labels. According to the results, both transfer across courses and in situ learning approaches have produced predictions that were actionable yet as accurate as those obtained with cross validation, suggesting that they deserve further attention to create impact in MOOCs with real-world interventions. Potential pedagogical uses of the predictions were illustrated with several examples.

Index Terms— engagement prediction, MOOC, peer review, transfer across courses, in situ learning

1 Introduction

MASSIVE open online courses (MOOCs) have fast increased in popularity over the past few years, enabling millions all around the world to receive free education in many subject areas with a basic Internet connection. Enrollments in MOOCs usually scale up to thousands [1] and contain a diverse population of learners with different knowledge levels, goals, learning preferences, and engagement styles [2]–[4]. While MOOCs are quite promising in democratizing education, its massive scale (alongside the variability among learners within the scale) leads to some pedagogical issues. This is largely because effective learning at large scale requires huge academic staff workload, which is not feasible to supply in practice. For example, MOOC instructors cannot possibly reply to every learner post in online discussions because participation in discussions at massive scale can easily lead to unmanageable overload of information [5]. Likewise, at large scales instructors cannot keep track of every learner, and therefore cannot provide timely formative feedback tailored to learners' distinct learning needs [6]. Learning analytics has been explored to offer mechanisms for providing personalized feedback in large courses [7]. Besides learning analytics, one common pedagogical solution to this issue has been the use of peer reviews [8].

Peer review is a reciprocal process, in which learners review the quality of peers' work while receiving a review from others on their own work [9]. Peer review has been already used extensively in classroom and online environments prior to MOOCs [10], [11], and the literature has been quite informative in demonstrating its learning benefits. Peer reviews have been often employed in MOOCs as an approach to assessing large numbers of learning actifacts [12]. However, its implementation at massive scale faces some challenges, limiting its benefits for learning. One important challenge is low participation in peer reviews [13], [14], which is not very surprising given the lack of instructor facilitation, diversity among MOOC learners, and high dropout rates. Low student participation can drastically hurt the effectiveness of peer reviews, resulting in a considerable number of student works with neither feedback to improve their work and learning nor a grade. Thus, Peer reviews bring certain capacities at large scales that are otherwise impractical. However, their true potential for supporting learning in MOOCs has not been well exploited.

As an attempt to contribute to the existing practice in MOOCs, the present study aims to propose a predictive approach to classify learners based on their expected level of peer-review engagement. In particular, we propose a classification approach to identify if a learner will under-participate in peer reviews or participate as required. Such categorization of learners can inform instructors' design decisions to mitigate issues related with peer-review implementation at massive scale. For example, some learners may under-participate because they might be actually lagging behind in the course and lack a solid understanding of concepts, and therefore, they may not feel confident in assessing others' work [14]. Such learners, if given additional time for performing peer reviews, can have the chance to catch up with others and improve their learning, which may lead to an increase in their participation in peer-reviews. Further, when knowing learners' expected level of participation, instructors can adjust the participation threshold for each learner separately (such as setting a higher threshold for learners who are expected to be highly engaged, and vice versa).

To utilize this approach in real-world practice, the predictions need to be available before the peer review activity takes place. In this concern, it is noticeable that most prediction research in the MOOC literature has instead focused on building models with post-hoc approaches in

which data from an already completed course (or activity) are used to train *and* validate predictive models [15], [16]. However, these models cannot be built until the target label is known (e.g., after learners had already dropped out), limiting their use in the very same courses from which they originate. The present research work is distinguished from prior research in that it proposes an approach to produce actionable predictions to be used in real-world practice. In particular, we use transfer across courses and in situ learning approaches, [17]–[19] to build a classification model. With both approaches, a prediction model is trained using only the information available at the time of prediction, and then used to make predictions early enough while they are still actionable for instructors' use. Our work is guided by the following research questions:

- To what extent can previous learner activities predict their engagement levels in peer reviews? What machine learning algorithms provide higher accuracy?
- Can the classification model that is trained on the preceding week's data perform well on the subsequent week (i.e., the performance of in situ learning)?
- Can the classification model that is trained using a completed course perform well on a new one (i.e., the performance of transferring across courses)?

The remainder of the paper is organized as follows. First, related work on peer reviews as well as on transfer across courses and in situ learning in the MOOC literature are presented. Then, the course datasets used to conduct this study are described, and their overall characteristics are discussed. Later, in the methods section, the features generated for building the classification models are described, and the approaches employed in building the models are presented. The last two sections present the results, discuss the findings, and suggest future research based on the limitations of the current work.

2 Related Work

2.1 Previous Research on Peer Reviews in MOOCs

The research on peer reviews in MOOC contexts has been diverse, however, heavily weighted towards the reliability and the validity of grading involved in peer reviews [8], [20]–[22]. Most of these studies have proposed mathematical models as an attempt to remove the noise in peer grading, resulted from the lack of expertise among assessors, and therefore improve its validity [8], [21]–[25]. Differently, the authors in [20] explored how different qualities of rubrics (used to guide the review process) could potentially improve the validity of peer grading. According to the findings, when the rubrics were improved with parallel sentence structures and unambiguous wording, an increase in agreement between peer-staff scores was observed, showing the importance of clear and well-communicated rubrics to achieve accurate peer grading. More recently in this category of research, the authors in [26] argued that peer assessment is a valid way to assess MOOC learners' performances as they found a significant correlation between scores of peer grading and learners' final exam scores.

Some other studies have explored the relationship between peer-review engagement and student learning. According to [12], peer-to-peer interactions through peer assessments enhance learners' understanding of concepts, and likewise the authors in [27], [28] found that engagement in peer-assessment tasks are strong determinants of learners' subsequent progress and performance. On the other hand, [14] reported that when learners have a good understanding of the concepts, then they are likely to not only review more peer work but also provide higher quality of feedback in their reviews. Thus, these studies suggest a reciprocal relationship: peer-review engagement enhances student learning, and at the same time, learners with a good understanding of concepts engage in peer-reviews with higher quantity and quality. Adding to these findings, the authors in [29] noted that high-achiever reviewers provide higher quality of feedback that help learners improve their performance significantly in the follow-up assignment. Moreover, some studies [14], [28] have investigated the influence of learners' demographic characteristics on their peer-review engagement and identified the geographic origin, employment status, education level, and weekly availability as the key factors. Last, in our previous work [13] we have focused on the prediction of the number of peer works that learners will review using machine learning regression algorithms. Although a certain degree of accuracy was obtained, the prediction models were trained using post-hoc methods, leading to minor practical use, and the data was highly imbalanced (in terms of class label distributions), creating bias in the predictions.

It is apparent that the literature thus far has made some significant contributions to peer-review practice in MOOCs. However, none of the previous studies have focused on proposing a solution to assist instructors in ongoing MOOCs in mitigating the problem of low learner engagement in peer reviews at large scales. Differently, and possibly complementary to the former studies, the current study aims to mitigate this issue by building a prediction model for early classification of learners based on their engagement-levels in peer reviews. As opposed to post-hoc prediction models, widely used in the MOOC literature, we propose a model that is operational for instructors' use in ongoing MOOCs. For this purpose, we take advantage of transfer across courses and in situ learning as described in the following section.

2.2 Transfer Across Courses and In-Situ Learning in MOOC Research

Most research on predictions within the context of MOOCs is devoted to dropout and performance prediction and uses data from a single past course to build and test predictive models with post-hoc approaches [30]. However, these approaches are not valid for real-world use since

training of the models requires labels that would not be available yet in real practice at the time when predictions are required by instructors [17]–[19], [30]. For example, the dropout prediction models proposed in [15], [16] are used to predict if a learner will drop out before the end of a course using models that were trained with labels that can only be obtained once the course is over. Thus, they could not be applied in real-world practice.

To overcome the limitations of post-hoc approaches, several works explored the use of the transfer across courses approach, in which a prediction model is built using a completed MOOC and then used for designing interventions in a follow-up MOOC [17], [31]. MOOCs themselves indeed offer distinct opportunities that make transfer learning an advantageous approach (e.g., transferring across re-runs of the same course, or across courses from the same domain or with similar instructional design) [17]. Nevertheless, there are not many studies that have investigated the potentials of transfer across MOOCs in comparison to post-hoc prediction approaches. Authors in [31] and [32] have tested the transferability of a dropout prediction model across different MOOCs. The results were quite promising, showing that different courses could be used to train a model to make predictions in another course. An increase in the accuracy of the predictions was noted when multiple courses were used to train the models, or when the training set was calibrated (e.g., maintain the learners in the training data that are more similar to the learners in the target course). Complementary to these findings, a recent study [19] has indicated that training a model on many other courses might lead to more accurate models compared to training on a course from the same discipline.

Different from transferring models across different MOOCs, the authors in [17] have proposed the in situ learning approach that allows training a model based on proxy labels (e.g., students are considered dropout if they have no interactions for a specific week [33]) and using this model when the prediction is needed for some intervention while the course is still continuing. A few studies have investigated the use of in situ learning in MOOCs. For example, in [18], [34], the researchers used in situ learning to predict if there will be a decrease in learner engagement at the end of a particular chapter (e.g., chapter 4) using the model trained on the preceding chapter data (e.g. chapter 3). Some other researchers [17], [19], [33] have tested in situ learning for building dropout prediction models that are transferable across different weeks within the same course and compared its performance with conventional transfer learning (using past courses). Both studies reported higher accuracy with in situ learning compared to transferring across courses. Thus, in situ learning could be a preferable technique over transferring between courses as it does not require different course data while producing accurate results.

Although, transfer across courses and in situ learning can provide actionable information for creating real-world interventions, their use is very limited in MOOC prediction research [30], and they have never been applied to peer review prediction. The current study uses both approaches to investigate their potentials for transferring prediction models for predicting peer-review engagement-level.

TABLE 1. SUMMARY OF THE COURSES IN TERMS OF ENROLLMENT, ASSIGNMENT, DISCUSSION, AND QUIZ ACTIVITIES

	Course#1	Course#2	Course#3
# of enrollments	3567	3632	5248
# of assignments (to be reviewed)	4	4	7
# of assignment submissions	2360	2015	1571
# of assignment submissions*	2488	2874	1671
# of peer reviews	5853	8006	2884
# of discussion topics	25	35	97
# of discussion entries	1249	2347	4852
# of quizzes	21	36	3
# of quiz submissions	14378	49837	1356
# of quiz submissions*	19253	82266	1419

* including learners' multiple submissions on the same assignment or quiz.

3 DATASETS

We have explored the Canvas Network repository shared in Harvard Dataverse[1] to determine proper courses to be used in the current research work. According to the Canvas Network Data Usage Agreement[2], the researchers of this work, who are the downloader of the data, are permitted to use this dataset for academic research and publication purposes. Three courses[3] were identified to carry out this research as they contain consistent data regarding peer reviews, and they were from the Business and Management domain as identified from the metadata about the course provided in the course data table.

All course data at hand were de-identified for privacy concerns by the publisher before their release. De-identification involved the removal of all textual information about the course (e.g., name of the discussion forums, assignment descriptions, discussion entries, etc.). The Canvas Network Data Usage Agreement prohibits to "produce connections or links among the information" and "extract information from the [dataset]". Therefore, no information regarding the learning design of the courses was known to the researchers. Nonetheless, in order to gain an overall insight into these courses, quantitative information regarding the number of enrollments as well as assignment, discussion, and quiz activities (which were the core components of three courses as inferred from the data at hand) were extracted. Summary of these information is provided in Table 1.

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XB2TLU.
[2] https://drive.google.com/file/d/0B5ghu5Vrh0j7YTlSdTRxVUg0Tnc/view

[3] Course IDs are: Course #1: 770000832960949, course #2: 770000832945397, and course #3: 770000832945322.

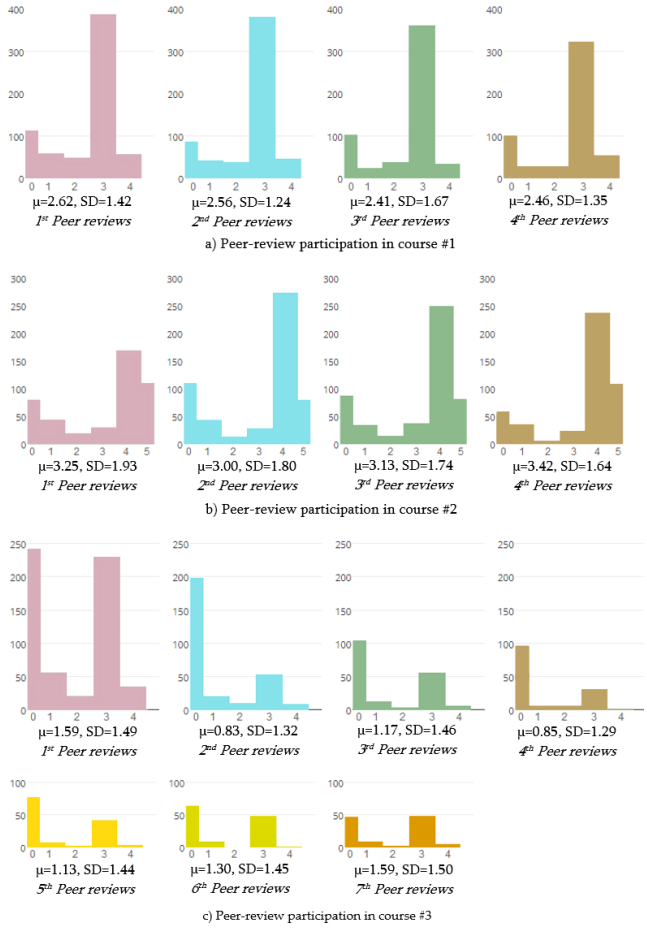

Figure 1. Learner participation in each peer-review session for (a) course #1, (b) course #2, and (c) course #3.

The basic statistics given in Table 1 may provide some useful insights into the courses. To begin with, course #1 and course #2 have exactly 4 assignments, and the number of submissions in these assignments are very close, which is parallel to the similarity between their enrollments. In comparison, course #3 contains more assignments yet less submissions in total even though it has the highest number of enrollments. These numbers may suggest that learner engagement in course #3 was problematic. Second, course #3 involved 7 assignments (and peer reviews) in comparison to course #1 and course #2 which involved four assignments in total to be reviewed by students. Therefore, it might have been more difficult to persist and complete all the activities in course #3. Moreover, compared to course #1, course #2 contains more discussion topics and quizzes, which is a probable reason for the higher number of discussion entries and quiz submissions. In particular, learners in course #2 seem to have an intense participation in quizzes. Also, learners in course #2 seem to be the most engaging group as they are the ones with the highest rates of assignment and quiz re-submissions. Furthermore, having 97 discussion topics and 4852 entries, the discussion forum seems to be a critical component of the learning design in course #3, whereas quizzes do not seem to be an integral component of learning: only three quizzes existed in this course.

In summary, course #1 and course #2 seem to have a similar design in comparison to course #3, which uses discussions heavily while having less focus on quizzes. Additionally, learner engagement in course #2 seems to be the highest whilst there is limited engagement in course #3. However, it is noteworthy that we cannot possibly have a 100% confidence on the design details of course components. For example, based on the data at hand, we cannot be sure about the types of the questions included in the quizzes (e.g., multiple-choice questions).

Besides the statistics mentioned above, we have examined learners' peer-review participation in each course, which is the variable of interest in the current study. There was one peer-review session per assignment in all courses (e.g., the first peer reviews for the first assignments, the second peer reviews for the second assignments, and so on). Using histograms, Figure 1 presents the overall participation in each peer-review session in their temporal order (from the first session till the last one). Mean and standard deviation scores are provided below histograms.

According to Figure 1, most of the learners in each course seem to review a certain number of peer works, which is 3 in course #1 and course #3, and 4 in course #2. These numbers are likely to be the level of participation expected from learners in the corresponding course, and as discussed previously, they will be used as thresholds to identify learners who are likely to under-participate in peer reviews versus learners who are likely to do as required. Moreover, learners' peer-review participation in all three courses seem to decrease as the semester progresses. The sharpest decrease was observed in course #3. One exception to this trend is that in course #2 the lowest participation in peer reviews was during the first session.

In course #1 and course #2, there was a clear separation among the time periods during which peer reviews for different assignments were performed. On the other hand, in course #3 there were many overlaps over these time periods due to the wide distributions of peer reviews over time. The peer views made after the due dates are regarded as no participation, since from instructors' point of view, knowing whether a learner will timely perform peer reviews might have more practical value than knowing whether a learner will eventually review peers' work. Dropping the late reviews has yielded many learners with zero participation in course #3 (see Figure 1), which was expected given the wide distribution of peer reviews in it.

The differences in the number of assignments, discussion topics, and quizzes, as well as their temporal distributions suggest that there is a considerably high probability that these courses are different and not re-runs of each other. This assumption neither can be entirely dismissed nor can be justified as no concrete information about the courses are released by the data provider.

4 METHOD

4.1 Feature Generation

Given that the current prediction task is about learners' peer-review participation, we focus on identifying aspects of learner activities in the course that could be somehow related to their peer-review participation. According to literature on peer reviews in MOOCs, there exists a strong relationship between learners' overall engagement and success in MOOCs and their peer-review participation [27], [28]. Therefore, we focus on generating features that are indicative of learners' performance and engagement in the core components of the courses, which are discussions, quizzes, assignments, and peer reviews. The list of the features generated are given in Table 2.

Most of these features (or very similar ones) have been already used in previous research as indicators of engagement [18], [35]. Indeed, in our preceding work [13] many of them were found to hold a predictive capacity on a similar task. Since the courses were completely de-identified, the learning design and the pedagogical intentions were unknown. As a result, such course information was not used to inform the feature selection. The features were computed using learner activity data accumulated starting from the first day of the course until the due date of the assignment for which learners' peer-review participation levels are predicted.

TABLE 2. FEATURES GENERATED FOR ALL COURSES

Discussion features	
x1: disc_count	Number of entries posted
x2: disc_charcount_mean	Average character length of entries posted
x3: disc_charcount_ttl	Total character length of entries posted
x4: disc_depth_mean	Average depth of entries posted
x5: disc_wordcount_mean	Average number of words in entries posted
x6: disc_wordcount_ttl	Total number of words in entries posted
Quiz features	
x7: finished_quiz_how_early	Number of days quiz was taken before the assignment due
x8: quiz_scores_mean	Average quiz score
x9: quiz_scores_ttl	Total quiz score
x10: quiz_timespent_mean	Average time spent on quizzes
x11: quiz_timespent_ttl	Total time spent on quizzes
x12: quiz_total_attempts_mean	Average quiz attempts
x13: quiz_total_attempts_ttl	Total quiz attempts
x14: uncomp_quiz_count	Number of incomplete quizzes
Assignment features	
x15: assign_attempt	Total number of assignment attempts
x16: assign_score	Assignment score
x17: assign_submt_how_early	Number of days assignment was submitted before its due
Peer-review features	
x18: pr_unique_count	Number of peer works reviewed
x19: message_size_ttl	Total size of feedback in bytes
x20: message_size_avg	Average size of feedback in bytes
x21: pr_timespent_ttl	Total time spent on peer reviews
x22: pr_timespent_avg	Average time spent on peer reviews
x23: message_size_multby _timespent	Message size in bytes multiplied by time spent on peer reviews
x24: pr_days_after_assig _subm	Number of days peer review was performed before assignment due

4.2 Prediction Models

4.2.1 Overall characteristics of the prediction models

Three popular machine learning algorithms, Logistic Regression, Random Forests, and Multi-Layer Neural Networks, ~~wereused~~ were used as the classifier. These algorithms have been effective in various classification tasks in the MOOC literature [18], [36]–[38]. The hyper parameters of estimators in all algorithms were tuned using (stratified 10-fold) cross-validated grid-search over a parameter grid. All features were standardized before they were used for training and testing the models. For training the models, we employed three different paradigms mentioned by Whitehill and his colleagues [33]: training on same course (post-hoc), training on other courses (transfer across courses), and training with proxy labels (in situ).

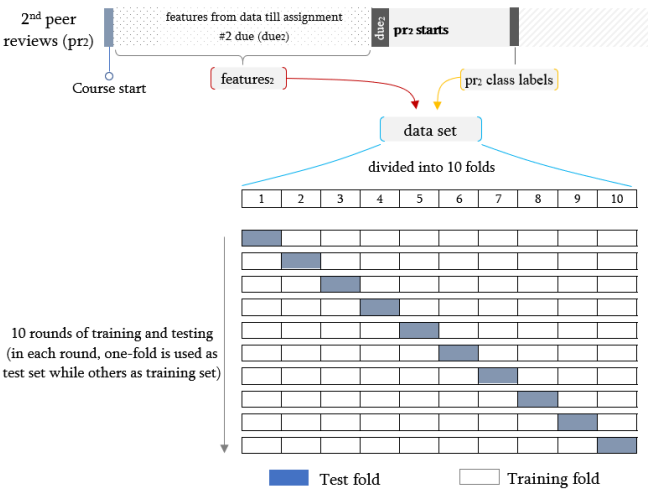

Figure 2. 10-fold CV to assess the model performance for the second peer reviews

4.2.2 Training on the same course (post-hoc)

This paradigm involves using the class labels from a completed course (or activity) to train a model first and then test its performance. This approach has no utility in an ongoing MOOC since the class labels (e.g., peer-review participation data) need to be known beforehand to train the models, which are however only available after the activity is completed (e.g., waiting until peer reviews are over). However, post-hoc models can be transferred to future courses for a possible real-world use (see section 4.2.3).

In this study, cross validation (CV) is used to evaluate the generalizability of post-hoc models to unseen data. In CV, the whole data is divided into equal chunks (called folds) [39]. Then, one of the folds is reserved for testing while the rest is used for training the model, and this process is repeated until each fold is once used as test set (e.g., 10-fold CV require 10 rounds of training and testing). We apply stratified 10-fold CV to each peer-review session in each course. In Figure 2, the use of 10-fold CV is illustrated for the second peer reviews.

Accuracy obtained with post-hoc training is likely to be

optimistic as a single dataset is used for training and testing [17]. However, this approach can serve as a point of reference to contrast with other training paradigms.

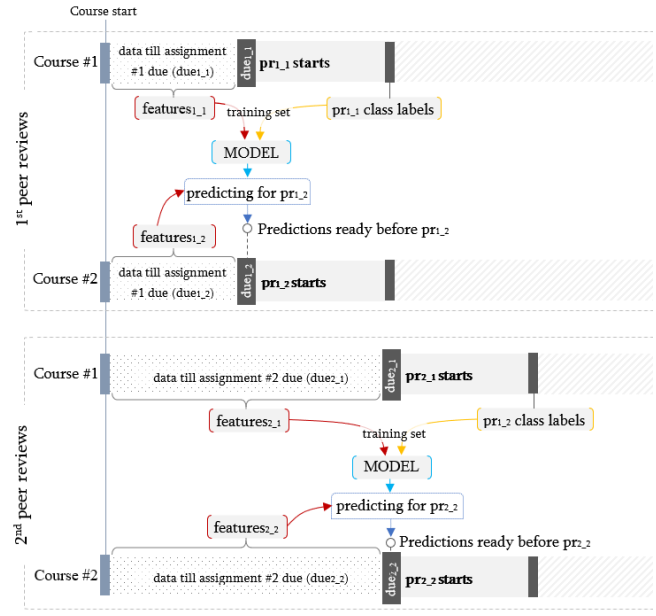

Figure 3. Transferring models per each peer-review session

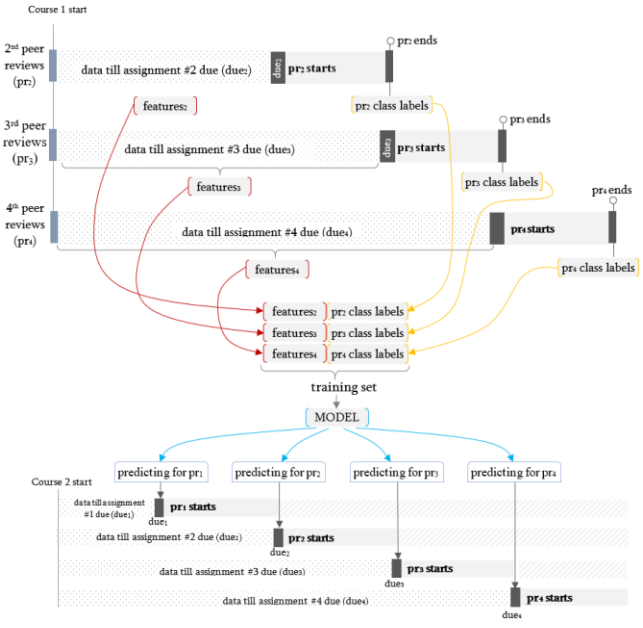

Figure 4. Transferring a single model across courses

4.2.3 Training on other courses (transfer across courses)

The current research aims to transfer models across courses and follows two specific methods for this purpose. The first method involves training a model per each peer-review session separately in one course and use each model to make a prediction in the corresponding peer-review session of another course. This approach obliges that the courses for transferring between should have a similar structure. Course #1 and course #2 indeed share structural similarity in the way peer-review sessions were organized: both courses contain four peer-review sessions with a one-week interval between each consecutive session. Therefore, we attempt to transfer models trained between course #1 and course #2 (and discarded course #3). To illustrate transferring per session between courses, Figure 3 depicts two example scenarios. In the first example, a model is built based on the data from the first peer reviews in course #1 (features1_1 and pr1_1 class labels) to predict learners' peer-review engagement levels on the due date of assignment #1 (due1_2) in course #2, just before the peer reviews (pr1_2) start. The second example is identical to the first one except that it is about predicting for the second peer reviews.

In the second method, (transferring per course), a single model is trained on the whole data from one course and this trained model is used to make separate predictions for each peer-review session in the other course instead of training separate models per each session. Data regarding the first sessions were excluded since they do not contain past peer-review activities, differently from the remaining sessions. Transferring per course is applied reciprocally among course #1, course #2, and course#3. Note that when using both transfer per session and full course transfer in real practice, the transfer can only happen from older courses (completed) to newer ones (ongoing). To illustrate transferring per course approach, Figure 4 depicts transferring a model from course #1 to course #2.

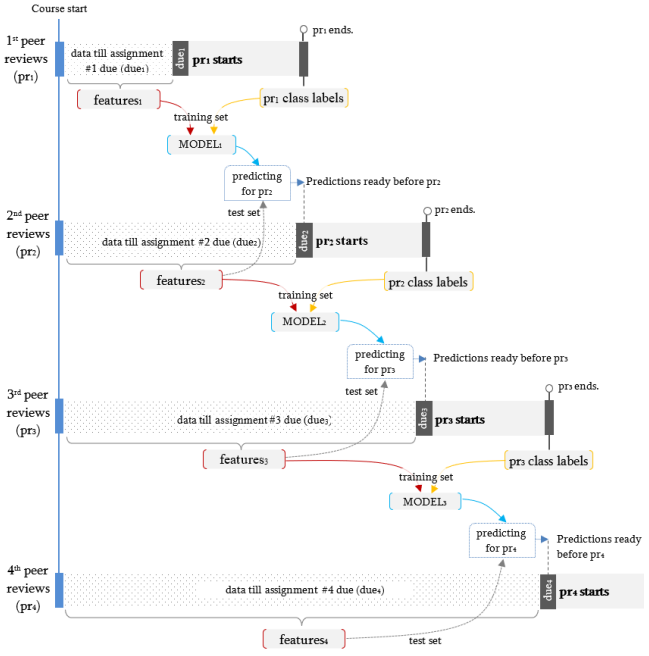

Figure 5. Transferring models using in situ learning in course #1 and course #2

Transferring a single model across courses is performed in a slightly different way as well: two of these three courses are combined to train a model (instead of using

only one course data), and this model was used for performing predictions in the other one. For example, a model is trained using the whole data from course #1 and course #2 combined, and this model is then used to make predictions per each peer-review session in course #3. This procedure was repeated for all combinations among three courses.

4.2.4 Training with proxy labels (in situ)

The last training approach is in situ learning, which allows transferring a model (trained possibly in the early phases of the course) over future prediction tasks within the same context. In our case, in situ learning involved the use of a model trained on an early peer-review data in a MOOC to classify learners by their peer-review engagement levels in a future peer-review session in the same MOOC.

With the use of in situ learning in the current task, the earliest prediction can be made regarding peer-review participation for the second assignment. This is because the data regarding the first peer-review session need to be used as the training set. The training set would be composed of (1) the class labels acquired from participation data in the first peer views, and (2) the features built on the learner activity data that emerged till the due date for the first assignment submissions. After being trained, the model is used to classify learners based on their expected level of engagement in reviewing peers' submissions for the second assignment. In the prediction, the test set was composed of the features calculated from the learner activity data that emerged until the assignment due. The same logic was applied to train and transfer models between two consecutive peer-review sessions (the former as the training set and the latter as the test set) in all three courses. Figure 5 demonstrates the use of in situ learning for course #1 and course #2. In the same manner, six models were trained in course #3 as there were seven peer-review sessions in total.

4.3 Assessment of Model Performance

Models were evaluated using area under the curve (AUC) as performance metric [40]. The AUC score of a classifier refers to the likelihood of ranking a randomly chosen positive example higher than a randomly chosen negative example [41]. A score of 0.5 would indicate a useless classifier (i.e., no different than a random classifier) whereas a score of 1.0 would indicate a perfect classifier (with 100% prediction accuracy). AUC, a commonly used metric in similar works in the literature [42], [43], was particularly chosen as it is considered rigorous to the prediction bias caused by imbalanced class distributions [33], which is the case in the current dataset (i.e., learners who under-participate in peer reviews are the minority in course #1 and course #2, whereas such learners are the majority in course #3).

In general, the categorization of model performance based on AUC scores follows as [44]–[46]: .9-1: excellent, .8-.9: very good, .7-.8: good, .6-.7: fair, and .5-.6: bad (or fail). In research on prediction of human behavior, AUC values of .7 and higher are considered reasonably accurate [47]. Similarly, previous MOOC research have considered models with such predictive power (i.e. AUC >= .7) robust [31]

5 Results

To compare the performances of all models, the accuracy scores are put together in a single table per each course and presented in Table 3 (for course #1), Table 4 (for course #2), and Table 5 (for course #3). Among the three algorithms (i.e., LR, RF, and NN), the one with the highest accuracy is indicated with bold font per each review session in each course. The results are presented per research question as follows.

TABLE 3. AUC SCORES OF PREDICTIONS IN COURSE #1 USING CV AND TRANSFER LEARNING APPROACHES

	1st reviews			2nd reviews			3rd reviews			4th reviews		
	LR	RF	NN	LR	RF	NN	LR	RF	NN	LR	RF	NN
CV	0.623	**0.624**	0.620	0.796	**0.806**	0.768	0.863	**0.874**	0.859	0.877	**0.892**	0.854
In situ	-	-	-	**0.617**	0.567	0.550	**0.863**	0.862	0.844	**0.857**	0.843	0.818
TpS#2	**0.588**	0.530	0.527	**0.769**	0.788	0.507	0.820	**0.838**	0.555	0.826	**0.845**	0.718
TpC#2	-	-	-	0.732	**0.768**	0.541	0.777	**0.794**	0.574	0.772	**0.814**	0.618
TpC#3	-	-	-	**0.760**	0.754	0.632	**0.793**	0.754	0.624	0.770	**0.772**	0.579
TpC#2&3	-	-	-	0.770	**0.803**	0.768	0.830	**0.835**	0.829	0.832	**0.859**	0.834

TpS#2: transferring per session from course #2, TpC#2: transferring per course from course #2, TpC#3: transfer per course from course #3, TpC#2&3: transferring per course from course #2 and course #3 combined.

TABLE 4. AUC SCORES OF PREDICTIONS IN COURSE #2 USING CV AND TRANSFER LEARNING APPROACHES

	1st reviews			2nd reviews			3rd reviews			4th reviews		
	LR	RF	NN	LR	RF	NN	LR	RF	NN	LR	RF	NN
CV	**0.606**	0.593	0.586	0.769	**0.772**	0.768	0.732	**0.784**	0.717	0.796	**0.801**	0.785
In situ	-	-	-	**0.689**	0.630	0.565	0.751	**0.759**	0.674	**0.779**	0.751	0.717
TpS#1	0.524	0.517	0.528	**0.707**	0.674	0.686	0.727	**0.734**	0.632	**0.798**	0.736	0.683
TpC#1	-	-	-	0.717	0.709	**0.722**	0.730	**0.735**	0.700	**0.813**	0.726	0.759
TpC#3	-	-	-	**0.665**	0.640	0.565	0.620	**0.676**	0.478	**0.715**	0.636	0.490
TpC#1&3	-	-	-	**0.719**	0.714	0.677	0.728	**0.754**	0.659	**0.812**	0.759	0.621

TpS#1: transferring per session from course #1, TpC#1: transferring per course from course #1, TpC#3: transfer per course from course #3, TpC#1&3: transferring per course from course #1 and course #3 combined.

TABLE 5. AUC SCORES OF PREDICTIONS IN COURSE #3 USING CV AND TRANSFER LEARNING APPROACHES

	1st reviews			2nd reviews			3rd reviews			4th reviews			5th reviews			6th reviews			7th reviews		
	LR	RF	NN	LR	RF	NN	LR	RF	NN	LR	RF	NN	LR	RF	NN	LR	RF	NN	LR	RF	NN
CV	0.561	**0.572**	0.546	0.712	0.868	0.892	0.743	0.830	**0.896**	0.757	0.830	**0.908**	0.832	0.832	**0.982**	0.852	0.818	**0.894**	0.824	0.843	**0.912**
In situ	-	-	-	**0.627**	0.607	0.475	**0.647**	0.641	0.539	0.722	**0.741**	0.704	**0.686**	0.684	0.606	0.717	**0.733**	0.688	0.730	**0.784**	0.683
TpC#1	-	-	-	**0.710**	0.677	0.573	**0.659**	0.600	0.604	**0.711**	0.604	0.606	0.670	**0.697**	0.578	**0.748**	0.718	0.672	0.681	**0.713**	0.585
TpC#2	-	-	-	0.643	**0.652**	0.630	0.547	**0.607**	0.508	0.590	**0.695**	0.544	0.640	**0.658**	0.591	0.663	**0.743**	0.554	0.589	**0.683**	0.504
TpC#1&2	-	-	-	**0.706**	0.608	0.674	**0.625**	0.615	0.526	**0.698**	0.587	0.560	**0.665**	0.659	0.654	**0.757**	0.736	0.650	0.721	**0.736**	0.592

TpS#1: transferring per session from course #1, TpC#1: transferring per course from course #1, TpC#2: transfer per course from course #2, TpC#1&2: transferring per course from course #1 and course #2 combined.

TABLE 6. FEATURE IMPORTANCE PER REVIEW SESSION IN EACH COURSE

	Course #1					Course #2					Course #3								Total (All)
	PR#1	PR#2	PR#3	PR#4	Total	PR#1	PR#2	PR#3	PR#4	Total	PR#1	PR#2	PR#3	PR#4	PR#5	PR#6	PR#7	Total	
pr_timespent_ttl	0	1	1	1	3	0	1	1	1	3	0	1	1	1	1	1	1	6	12
pr_unique_count	0	1	1	1	3	0	1	1	1	3	0	1	0	1	1	1	1	5	11
pr_message_size_multby_timespent	0	1	1	1	3	0	1	1	1	3	0	0	1	1	1	1	1	5	11
pr_msg_size_ttl	0	1	1	1	3	0	1	1	1	3	0	0	1	1	0	1	1	4	10
pr_subs_count	0	1	1	1	3	0	1	1	1	3	0	1	0	1	0	1	1	4	10
pr_days_after_assig_subm	0	1	1	1	3	0	0	1	1	2	0	1	0	0	1	1	1	4	9
pr_message_size_avg	0	1	1	1	3	0	1	1	1	3	0	0	0	0	1	0	1	2	8
assign_submt_how_early	0	1	1	1	3	0	0	1	1	2	0	0	1	0	0	1	1	3	8
pr_timespent_avg	0	0	1	1	2	0	1	1	1	3	0	1	1	0	0	1	0	3	8
finished_quiz_how_early	1	1	0	1	3	0	0	0	0	0	1	0	1	1	1	0	0	4	7
quiz_time_spent_ttl	1	1	0	0	2	0	0	0	0	0	1	1	1	1	1	0	0	5	7
disc_wordcount_ttl	1	0	0	0	1	1	0	0	0	1	1	1	1	1	0	0	0	4	6
quiz_time_spent_mean	1	0	0	0	1	1	0	0	0	1	1	1	1	0	1	0	0	4	6
assign_score	0	0	0	0	0	0	1	0	1	2	0	0	0	1	0	1	1	3	5
disc_charcount_ttl	1	0	1	0	2	1	0	0	0	1	1	1	0	0	0	0	0	2	5
disc_count	0	0	0	0	0	1	0	0	0	1	0	1	1	1	1	0	0	4	5
disc_depth_mean	0	0	0	0	0	1	0	0	0	1	1	0	0	0	1	1	1	4	5
quiz_scores_ttl	1	0	0	0	1	1	1	1	0	3	0	0	0	0	0	0	0	0	4
disc_charcount_mean	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	1	3
quiz_total_attempts_mean	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	1	3
quiz_total_attempts_ttl	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	1	3
disc_wordcount_mean	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	2
quiz_scores_mean	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	2

5.1 RQ1: The predictive power of past course activities and the performance of different machine learning algorithms

Regarding the performances achieved with different algorithms, in most cases LR and RF produced more accurate results when compared to NN in all courses. Although LR and RF have performed very similarly, the accuracy achieved with RF was relatively higher in many cases.

A Wilcoxson signed rank test was conducted to identify statistically significant differences among the algorithms. The test results showed that both LR and RF significantly performed better than NN ($p < 0.05$) whereas there was not a significant performance difference between LR and RF ($p=0.640$). These findings were consistent across three courses.

In order to identify the predictive power of the features, feature importance was calculated for the models that were built with RF, which produced the most accurate scores in overall. The importance of the features is automatically computed in RF implementation of Scikit-Learn [48]. The most important 10 features identified in each model of peer-review prediction (in all courses) are merged into a single list as shown in Table 6, leading to ~~22~~ 23 features in total. These features are ordered by how often they appeared in the top ten list across all models. According to results, features derived from students' past peer-review engagement data, corresponding to 9 features, were the most predictive consistently in all courses and in all peer review sessions (except the first ones without a past peer review activity). Among those, the total amount of time spent on peer review was the most important feature.

Moreover, two assignment-related features were found to be predictive in many models (particularly, in the first two courses). Almost all quiz-related features were predictive, especially when building the models for the first peer reviews, in which features derived from past peer-review activities was not available. In particular, how much time students spent on the quizzes and how soon (or late) they took the quiz before the peer-reviewed assignment were

more significant. Among the three courses, quiz-related features were barely predictive in the second course. Last, discussion-related features were found to be predictive in relatively less models. These features were more significant in the first peer-review sessions.

5.2 RQ2: The performance of in situ learning approach

In situ learning technique was applied to predict engagement-levels in peer reviews using the models trained on preceding peer-review activity data in the three. This technique was used starting from the second review sessions (i.e., the earliest session with past peer-review data available).

The in situ learning approach yielded models mostly with good and very good accuracy levels (particularly in later predictions) and some with fair level of accuracy (particularly in earlier predictions). Predictions were the least accurate for the 2nd peer reviews in all courses, where the models did not include the features about past peer-review activities. The most accurate predictions were achieved in the first course, while the least accurate ones were in the third course. In all courses, the accuracy increased across peer-review sessions. LR and RF were the best performing algorithms in all cases.in situin situin situ

5.3 RQ3: The performance of transferring models across courses

Models were transferred across courses in two manners. First, regarding the transferring per session, accuracies were the lowest when transferring the models between the first peer-review sessions across courses. However, the predictions were accurate starting from the second peer reviews with an increasing performance in each subsequent review session. In general, the models performed generally very good in course #1 and good in course #2 (except for the first peer reviews in both courses). Comparing with the in situ approach, transferring per session produced slightly more accurate predictions in the second peer-review sessions, whereas in the later sessions the predictions with in situ approach were slightly more accurate. These results were observed in both courses #1 and #2.

Second, regarding the transfer of models trained on an entire course to other courses, the accuracies were not consistent. The transfers between course #1 and course #2 yielded good or very good accuracy levels (with model transferred from course #2 being slightly more accurate). The transfer from course #3 to course #1 produced accurate predictions that are however slightly lower than those obtained with transferring from course #2. The results were very similar when the transfer was from course #3 to course #2 (i.e., accuracies were good in general but lower than those obtained with transferring from course #1). In all courses, the transferred models performed slightly better in each subsequent peer-review session.

Moreover, considering the cases where the transfer was based on two courses combined (e.g., training a model on course #1 and course #2 data together, and transferring it to course #3) the accuracies of predictions were quite similar to those obtained with models trained on one course dataset.

6 Discussion

The discussion of the results is presented per each research question separately as follows.

6.1 RQ1: The predictive power of past course activities and the performance of different machine learning algorithms

According to the analysis of feature importance, almost all features included in the models, (i.e., ~~21~~ 23 (see Table 6) out of 24 (see Table 2) demonstrated some predictive power. That is, the features used to build the prediction models are highly relevant to the current classification task. Unsurprisingly, past peer-review activities were the most powerful to predict the future peer-review engagement. This finding is consistent with the literature reporting high predictive power of engagement indicators of past activities (e.g., submission history) that are directly related to the target activity (final exam or dropout) [35]. Therefore, the features about past peer-review activities were the ones with the highest capacity to generalize across peer-review sessions and across courses.

Although they were predictive in the first peer reviews, the importance of the rest of the features was rather inconsistent across courses. For example, quiz scores did not demonstrate a predictive capacity in the third course, while they were more significant in the second course. Similarly, number of discussion counts did not matter in the first course, whereas they were predictive in several models in the third course. We argue that this inconsistency is associated with the different learning design applied in each course [49]. That is, for example, the way discussion forums are used pedagogically and connected with the peer-reviewed activity may differ from one course to another, which may create a weaker/stronger association between student engagement in discussions and the peer reviews.

Regarding the poor performance of NN, relatively low sample size might have played a role. The reason why LR produced accurate results could be that there was a rather simple linear relationship between the features and the target variable. It is noteworthy that at each subsequent prediction, the AUC scores increased with each algorithm, which is probably due to the noise caused by samplers and strong starters [50]. That is, there were quite many learners early in the courses who were exploring the course pages with inconsistent behavior and most of these learners probably dropped out in the following weeks.

6.2 RQ2: The performance of in situ learning approach

According to the results regarding in situ learning, the model built on activities prior to the current peer-review

session performs almost as good as the model that could be eventually built once the peer-review session is over (which is only feasible in a post-hoc analysis such as CV, but not in real practice). The high accuracy achieved with in situ learning is in congruence with the results regarding feature importance. That is, the predictive power of past peer-review activities was an indicator of their capacity to be used as a proxy label to train models with in situ learning.

The advantage of in situ over transferring between courses is that it does not require any data from past MOOCs, and it can be still used in an ongoing MOOC to generate operational predictions. However, this approach is limited in terms of predicting the participation in the first peer-review session since it is mandatory to have a past peer-review data for training a model. Therefore, no prediction accuracy was reported for the first peer reviews when using in situ learning. Another disadvantage is that the models trained on the data from the first peer-review sessions were lacking features regarding learners' past peer-review activities, therefore they had limited capacity in predicting participation in the second sessions. These limitations were less adverse in course #3, where there were relatively higher number of peer-review sessions.

6.3 RQ3: The performance of transferring models across courses

According to the results, the models transferred between the matching peer-review sessions of course #1 and #2 were able to accurately predict student engagement in peer reviews. However, the transfer between the first sessions of these courses, in which the model lacked the peer-review features, was not successful.

Transferring per session produced more accurate results in the second peer reviews in comparison to those with in situ learning. This was probably because the models transferred included the features about past peer-review activities whereas those trained with in situ learning on the first peer-review activity data lacked these features. Thus, the technique of transferring the models per session across courses can be more advantagous than in situ learning when predicting engagement in early activities since it allows to incorporate more features in the model.

The transfer of the models trained on the whole course data yielded accurate predictions. However, the accuracy was slightly different depending on the courses. For example, a higher predictive power with transfers between course #1 and course #2 versus course #3 was achieved. This result might be due to the fact that course #1 and course #2 have a similar course structure, and therefore they might share some commonalities in their learning designs, whereas course #3 was identified to have rather different design. Moreover, considering the cases where the transfer was based on two courses combined (e.g., training a model on course #1 and course #2 data together, and transferring it to course #3) the accuracies of predictions were quite similar to those obtained with models trained on one course dataset. As derived from the comprehensive analysis of the available anonymized data (see section 6 Dataset for a detailed discussion), these are different courses (i.e., not replication of the same course). These results suggest that there is a great potential to utilize for transferring predictive models across different MOOC contexts. Previous research has noted similar findings when models derived from different MOOCs are used for dropout prediction in another MOOC [19].

7 Implications

7.1 Implications for Pedagogy

The predictions produced by the classification models presented in this work could be used in numerous ways to support the pedagogy of an ongoing MOOC. However, only the course instructors could precisely identify the uses of these predictions based on the contextual factors and pedagogical needs. The decision on how to use the predictions could depend on the perspective and goals of the instructor, who may opt to personalize the learning path of the learners [51] or to gamify some learners' learning experience to increase their participation in some activities [52], [53] or who may decide to take actions that favor the global performance without differentiating student groups. Nonetheless, we provide several ideas to demonstrate the potential uses of the predictions in practice. For example, based on learners' predicted level of peer-review participation, the match between peers could be improved to ensure that every learner receives a feedback on their work, which is crucial to their learning process [54]. A mathematical model could be developed based on each learner's probability of performing a peer review (ranging between 0.0-1.0), and this model would allow instructor to determine the minimum probability (*PR*) that a learner submission will receive at least a peer review. This model could be simply based on the following formula:

$$PR = 1 - (1 - p_1) \times (1 - p_2) \times \ldots \times (1 - p_n) = 1 - \prod_{i=0}^{n} (1 - p_i)$$

Where p is the probability that a learner will performe a peer review, and n is the number of learners assigned to calculate the PR of the intended peer-review session.

For example, the instructor may require that each submission should be reviewed at least by 1 peer with a probability of 95%. In this case, this model could be employed to automatically identify n peers with the desired p values so that the 95% of *PR* value could be obtained. The instructor should be aware that 1-*PR* (i.e., 5%) of submissions may not receive any reviews, and they would need to be addressed somehow (e.g., assigning them to learners who have already done at least 1 review).

The proposal above illustrates only one potential use of the information provided by the predictor, but many other uses could be envisioned, depending on the issues that concern the instructor the most. For example, if the predictions suggest a critically low participation in peer reviews, the instructor could attempt to identify the reasons for low interest in peer reviews, and then take some actions to increase learners'

motivation in reviewing a peer work (like making them mandatory to obtain the certification or introducing gamification incentives). Besides the potential of the proposed classification approach to improve the existing peer-review practice in MOOCs, it can also help improve the learning design of MOOCs depending on the objective of the intervention as well as the transferring approach used. For example, with the goal of improving the learning design, an instructor may decide to change a particular assignment completely and add some exercises prior to the assignment in the re-run of a previous MOOC. Whether these changes will have the desired influence (or not) can be identified with the help of a classification model trained on the data from the initial run of the course. The instructor can obtain learners' expected level of participation before the actual peer reviews start, make a comparison, and decide any further changes are needed. Similarly, transferring over weeks of the same course might offer such benefits. For example, predictions indicating low-level of peer-review participation may actually suggest some problems rooted in the learning design of the current learning module. After revisiting the module (e.g., reviewing the examples provided, reading learners' discussions), the instructor may identify some issues (lack of supplementary learning materials) and resolve them.

7.2 Implications for Theory and Design

In the current study, the models transferred between MOOCs with relatively similar learning design led to more accurate results. One theoretical implication of this finding is that the performance of machine learning models transferred across different MOOCs is closely associated with the match between the learning designs. If this match is stronger, then the models transferred produce more accurate predictions. This is about the way learning design shapes how students may engage and interact with different components of a course. For instance, a discussion forum may involve instructor to play a central role or to act as a facilitator to promote peer interactions depending on the pedagogical intentions, which may result in distinct student behavior (and engagement data) in both cases [55]. However, if different courses adopt a similar pedagogical intention in their discussion forums, the way students engage in them might be alike and may lead to engagement indicators with similar predictive capacity. That is, learning design is an important factor to consider when transferring predictive models across different courses.

A related implication is about the way open data is published. The data used in this study was lacking details about the learning designs of the course. This issue of lacking context in educational data draws a particular attention to current practice in sharing educational data publicly. As noted in the current study, research on educational data that lacks the pedagogical context is likely to lead to relatively speculative findings from which the learning design is detached. Therefore, decision-makers on data sharing should consider including information regarding learning design and context. In line with this argument, there have been growing efforts toward open science and data sharing [56].

Given the role of learning design, another important implication is that instructors might be involved in the design process of predictive models. With instructor input, a richer set of features could be generated that may have a higher predictive power. Collaboration with instructors may also produce a predictive analytics solution that aligns better with contextual needs. Additionally, when involved in the process, instructors may be inclined to use the actionable models to improve their teaching practice. Several research highlighted the benefits emerging from involving instructors in the loop of designing predictive models [35], [49].

8 Conclusions and Future Work

This study has presented a novel approach for classifying-MOOC learners based on their expected level of engagement in peer reviews. The results have shown that with transfer across courses and in situ learning approaches in the current classification task can produce predictions that are accurate and actionable for using in real practice. Therefore, this approach can be incorporated into real MOOC contexts to produce actionable information (e.g., classification of learners based on their expected level of participation in peer reviews) at the time when instructors need it and still can use it. In this way, instructors can take some remedial actions to mitigate emerging problems and also improve the learning design of their courses. For example, they can use these predictions to improve participation in peer reviews or design effective collaborative learning activities.

The approaches used in this work can be implemented in other MOOC contexts for different prediction tasks. For example, in a MOOC where there exist several exams (or quizzes), in situ approach could be utilized to predict in real time if learners will pass or fail an upcoming exam. Similarly, a model could be trained from another MOOC (with similar course design) and transferred to the new context to train a model for such a prediction. In Figure 3-5, we have provided illustration regarding the application of these machine learning techniques to guide the future researchers in employing them in their own studies.

However, one limitation of this work is that the research was conducted in decontextualized MOOCs; therefore, the predictive models lack the understanding of the learning design and the context. As a follow-up research, we plan to study the proposed classification approach in courses where the pedagogical context is known so that the learning design and the predictive analytics are informed by each other. Related to this follow-up work, we also plan to collaborate with MOOC instructors to explore the possible uses of the actionable information offered by the prediction approach and study its impact. Furthermore, as noted in previous research [28], [37], demographic information about learners can be highly associated with learners' peer-review engagement; however, such information was missing in the current study. Therefore, in the next phase of this research line, we plan to administer a survey to collect some demographic information from learners (e.g., employment status, education level, etc.) and create

additional features from this information to improve the prediction performance. Moreover, the current study can motivate some future research on collaborative learning at large scales. In this regard, we plan to integrate the current work with previous group formation approaches [57], and examine its effectiveness in building collaborative teams. Last, we have used AUC as the performance metric to overcome the bias introduced by imbalanced data. Future research should explore other approaches such as subsampling to overcome problems about imbalanced data.

Acknowledgments

This research has been fully funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement 793317, and partially funded by the European Regional Development Fund and the National Research Agency of the Spanish Ministry of Science, Innovations and Universities under project grants TIN2017-85179-C3-2-R and TIN2014-53199-C3-2-R, by the European Regional Development Fund and the Regional Ministry of Education of Castile and Leon under project grant VA257P18, by the European Commission under project grant 588438-EPP-1-2017-1-EL-EPPKA2-KA, and. Access to the data used in this paper was granted by Canvas Network.

References

[1] S. Rayyan *et al.*, "A MOOC based on blended pedagogy," *J. Comput. Assist. Learn.*, vol. 32, no. 3, pp. 190–201, 2016.

[2] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses," in *Proceedings of Third International Conference on Learning Analytics and Knowledge*, 2013, pp. 170–179.

[3] T. Phan, S. G. Mcneil, and B. R. Robin, "Students' patterns of engagement and course performance in a Massive Open Online Course," *Comput. Educ.*, vol. 95, pp. 36–44, 2016.

[4] C. Alario-Hoyos, M. Perez-Sanagustin, C. Delgado-Kloos, H. A. P. G., and M. Munoz-Organero, "Delving into participants' profiles and use of social tools in MOOCs," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, 2014.

[5] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong, "Learning about social learning in MOOCs: From statistical analysis to generative model," *IEEE Trans. Learn. Technol.*, vol. 7, no. 4, pp. 346–359, 2014.

[6] H. Suen, "Peer assessment for massive open online courses (MOOCs)," *Int. Rev. Res. Open Distrib. Learn.*, vol. 15, no. 3, 2014.

[7] L.-A. Lim *et al.*, "What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course," *Learn. Instr.*, vol. In Press, 2019.

[8] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proceedings of Sixth International Conference on Educational Data Mining*, 2013, pp. 153–160.

[9] K. Topping, "Peer assessment between students in colleges and universities," *Rev. Educ. Res.*, vol. 68, no. 3, pp. 249–276, 1998.

[10] E. F. Gehringer, "Electronic peer review and peer grading in computer-science courses," *ACM SIGCSE Bull.*, vol. 33, no. 1, pp. 139–143, 2001.

[11] D. Nicol, A. Thomson, and C. Breslin, "Rethinking feedback practices in higher education: a peer review perspective," *Assess. Eval. High. Educ.*, vol. 39, no. 1, pp. 102–122, 2014.

[12] D. K. Comer, C. R. Clark, and D. A. Canelas, "Writing to learn and learning to write across the disciplines: Peer-to-peer writing in introductory-level MOOCs," *Int. Rev. Res. Open Distance Learn.*, vol. 15, no. 5, pp. 26–82, 2014.

[13] E. Er, M. L. Bote-Lorenzo, E. Gómez-Sánchez, Y. Dimitriadis, and J. I. Asensio-Pérez, "Predicting student participation in peer reviews in MOOCs," in *Proceedings of the Second European MOOCs Stakeholder Summit 2017*, 2017.

[14] S. Meek, L. Blakemore, and L. Marks, "Is peer review an appropriate form of assessment in a MOOC? Student participation and performance in formative peer review," *Assess. Eval. High. Educ.*, pp. 1–14, Aug. 2016.

[15] S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, "Predicting MOOC performance with week 1 behavior," in *Proceedings of the 7th International Conference on Educational Data Mining*, 2014, pp. 273–275.

[16] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization," *Comput. Human Behav.*, vol. 58, pp. 119–129, 2016.

[17] S. Boyer and K. Veeramachaneni, "Transfer learning for predictive models in Massive Open Online Courses," in *Proceedings of the 17th Conference on Artificial Intelligence in Education*, Madrid, Spain, 2015, pp. 54–63.

[18] M. L. Bote-Lorenzo and E. Gómez-Sánchez, "Predicting the decrease of engagement indicators in a MOOC," in *Proceedings of Seventh International Conference on Learning Analytics and Knowledge*, 2017, pp. 143–147.

[19] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, "MOOC dropout prediction: How to measure accuracy?," in *Proceedings of the Fourth ACM Conference on Learning@Scale*, 2017, pp. 161–164.

[20] C. Kulkarni *et al.*, "Peer and self assessment in massive online classes," *ACM Trans. Comput. Interact.*, vol. 20, no. 6, pp. 1–31, 2013.

[21] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran, "A Case for Ordinal Peer-evaluation in MOOCs," in *NIPS Workshop on Data Driven Education*, 2013, pp. 1–8.

[22] J. Diez, O. Luaces, A. Alonso-Betanzos, A. Troncoso, and A. Bahamonde, "Peer Assessment in MOOCs Using Preference Learning via Matrix Factorization," *NIPS Work. Data-Driven Educ.*, no. DECEMBER, pp. 1–6, 2013.

[23] O. Luaces, J. Díez, A. Alonso-Betanzos, A. Troncoso, and A. Bahamonde, "Content-based methods in peer assessment of open-response questions to grade students as authors and as

graders," *Knowledge-Based Syst.*, vol. 117, pp. 79–87, 2017.

[24] N. Capuano and S. Caballé, "Towards adaptive peer assessment for MOOCs," in *Proceedings of 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 2015, pp. 64–69.

[25] N. Capuano, V. Loia, and F. Orciuoli, "A fuzzy group decision making model for ordinal peer assessment," *IEEE Trans. Learn. Technol.*, vol. 10, no. 2, pp. 247–259, 2017.

[26] W. Admiraal, B. Huisman, and M. Van De Ven, "Self- and peer assessment in Massive Open Online Courses," *Int. J. High. Educ.*, vol. 3, no. 3, pp. 119–128, 2014.

[27] M. M. Ashenafi, M. Ronchetti, and G. Riccardi, "Predicting student progress from peer-assessment data," in *Proceedings of the 9th International Conference on Educational Data Mining*, 2016, pp. 270–275.

[28] M. Formanek, M. C. Wenger, S. R. Buxner, C. D. Impey, and T. Sonam, "Insights about large-scale online peer assessment from an analysis of an astronomy MOOC," *Comput. Educ.*, vol. 113, pp. 243–262, 2017.

[29] B. Huisman, W. Admiraal, O. Pilli, M. van de Ven, and N. Saab, "Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance," *Br. J. Educ. Technol.*, 2016.

[30] J. Gardner and C. Brooks, "Student success prediction in MOOCs," *User Model. User-adapt. Interact.*, vol. 28, no. 2, pp. 127–203, 2018.

[31] S. Boyer and K. Veeramachaneni, "Robust predictive models on MOOCs: Transferring knowledge across courses," in *Proceedings of the Ninth International Conference on Educational Data Mining*, 2016, pp. 298–305.

[32] S. Boyer, B. U. Gelman, B. Schreck, and K. Veeramachaneni, "Data science foundry for MOOCs," in *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*, 2015, pp. 1–10.

[33] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, "Delving deeper into MOOC student dropout prediction," *arXiv*, pp. 21–27, 2017.

[34] M. L. Bote-Lorenzo and E. Gómez-Sánchez, "An approach to build in situ models for the prediction of the decrease of academic engagement indicators in massive open online courses," *J. Univers. Comput. Sci.*, vol. 24, no. 8, pp. 1052–1071, 2018.

[35] K. Veeramachaneni, U.-M. O'Reilly, and C. Taylor, "Towards feature engineering at scale for data from massive open online courses," *arXiv*, 2014.

[36] R. F. Kizilcec, M. Pérez-Sanagustín, and J. J. Maldonado, "Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses," *Comput. Educ.*, vol. 104, pp. 18–33, 2016.

[37] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A review and future research directions," *IEEE Trans. Learn. Technol.*, no. In Press, 2018.

[38] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, and C. D. Kloos, "Analysing the predictive power for anticipating assignment grades in a massive open online course," *Behav. Inf. Technol.*, vol. 37, no. 10–11, pp. 1021–1036, 2018.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. 2009.

[40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[42] M. Mouriño-García, R. Pérez-Rodríguez, L. Anido-Rifón, M. J. Fernández-Iglesias, and V. M. Darriba-Bilbao, "Cross-repository aggregation of educational resources," *Comput. Educ.*, vol. 117, no. September 2017, pp. 31–49, 2018.

[43] C. D. B. Luft, J. S. Gomes, D. Priori, and E. Takase, "Using online cognitive tasks to predict mathematics low school achievement," *Comput. Educ.*, vol. 67, pp. 219–228, 2013.

[44] A.-M. Šimundic, "Measures of diagnostic accuracy: basic definitions," *EJIFCC*, vol. 19, no. 4, pp. 203–211, 2009.

[45] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–39, 2013.

[46] T. G. Tape, "The area under an ROC curve.," 2008. [Online]. Available: http://gim.unmc.edu/dxtests/roc3.htm. [Accessed: 22-Jul-2017].

[47] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r," *Law Hum. Behav.*, vol. 29, no. 5, pp. 615–620, 2005.

[48] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," vol. 12, pp. 2825–2830, 2012.

[49] E. Er, E. Gómez-Sánchez, Y. Dimitriadis, M. L. Bote-Lorenzo, J. I. Asensio-Pérez, and S. Álvarez-Álvarez, "Aligning learning design and learning analytics through instructor involvement: A MOOC case study," *Interact. Learn. Environ.*, vol. 27, no. 5–6, pp. 685–698, 2019.

[50] R. Ferguson and D. Clow, "Examining engagement : Analysing learner subpopulations in Massive Open Online Courses (MOOCs)," in *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*, 2015, pp. 51–58.

[51] D. Ifenthaler and D. Eseryel, *Adapting for a Personalized Learning Experience*. 2013.

[52] L. Ding, E. Er, and M. Orey, "An exploratory study of student engagement in gamified online discussions," *Comput. Educ.*, vol. 120, 2018.

[53] A. Ortega-Arranz, J. A. Muñoz-Cristobal, A. Martínez-Monés, M. L. Bote-Lorenzo, and J. I. Asensio-Pérez, "How Gamification is Being Implemented in MOOCs? A Systematic Literature Review," in *Proceedings of the 12th European Conference on Technology Enhanced Learning*, 2017, vol. 10474, pp. 441–447.

[54] J. Hattie and H. Timperley, "The power of feedback," *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007.

[55] L. Lockyer, E. Heathcote, and S. Dawson, "Informing

pedagogical action: aligning learning analytics with learning design," *Am. Behav. Sci.*, vol. 57, no. 10, pp. 1439–1459, 2013.

[56] B. A. Nosek *et al.*, "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility," *Science*, vol. 348, no. 6242, pp. 1422–1425, 2015.

[57] L. Sanz-Martínez, J. A. Muñoz-Cristobal, M. L. Bote-Lorenzo, A. Martínez-Monés, and Y. Dimitriadis, "Toward criteria-based automatic group formation in MOOCs," in *Proceedings of the 5th European MOOCs Stakeholders Summit*, 2017.