Seeding Course Forums using the Teacher-in-the-Loop

Einat Shusterman
Ben-Gurion University, Israel
einatshu@post.bgu.ac.il

Hyunsoo Gloria Kim
Marc Facciotti
Michele Igo
Kamali Sripathi
University of CA, Davis, USA
{hyunsookim,mtfacciotti,mmIgo,ksripathi}@ucdavis.edu

David Karger
MIT
karger@mit.edu

Avi Segal
Kobi Gal*
Ben-Gurion University, Israel
{avise,kobig}@bgu.ac.il

ABSTRACT

Online forums are an integral part of modern day courses, but motivating students to participate in educationally beneficial discussions can be challenging. Our proposed solution is to initialize (or "seed") a new course forum with comments from past instances of the same course that are intended to trigger discussion that is beneficial to learning. In this work, we develop methods for selecting high-quality seeds and evaluate their impact over one course instance of a 186-student biology class. We designed a scale for measuring the "seeding suitability" score of a given thread (an opening comment and its ensuing discussion). We then constructed a supervised machine learning (ML) model for predicting the seeding suitability score of a given thread. This model was evaluated in two ways: first, by comparing its performance to the expert opinion of the course instructors on test/holdout data; and second, by embedding it in a live course, where it was actively used to facilitate seeding by the course instructors. For each reading assignment in the course, we presented a ranked list of seeding recommendations to the course instructors, who could review the list and filter out seeds with inconsistent or malformed content. We then ran a randomized controlled study, in which one group of students was shown seeds that were recommended by the ML model, and another group was shown seeds that were recommended by an alternative model that ranked seeds purely by the length of discussion that was generated in previous course instances. We found that the group of students that received posts from either seeding model generated more discussion than a control group in the course that did not get seeded posts. Furthermore, students who received seeds selected by the ML-based model showed higher levels of engagement, as well as greater learning gains, than those who received seeds ranked by length of discussion.

ACM Reference Format:
Einat Shusterman, Hyunsoo Gloria Kim, Marc Facciotti, Michele Igo, Kamali Sripathi, David Karger, Avi Segal, and Kobi Gal. 2021. Seeding Course Forums using the Teacher-in-the-Loop. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3448139.3448142

1 INTRODUCTION

Discussion forums have been used successfully as tools to facilitate interactions and exchanges of knowledge between learners and between learners and instructors (Poole, 2000). Promoting student discussion in online forums has been linked to increased learning gains (Cormier and Siemens, 2010, DeSanctis et al., 2003). Online discussion forums have also been shown to promote an increased exchange of ideas, as well as improve students' ability to make connections between concepts and to apply the course material to diverse contexts (Breslow et al., 2013).

However, not all forms of discussion are beneficial for learning (Romeo, 2001). Online discussions that do not promote higher levels of thinking are ineffectual in providing increased learning (Wang et al., 2015). Studies have also shown that pedagogical benefit arises only when discussion encourages students to share different interpretations and perspectives of the course (Light et al., 2000).

In this work, we study how to improve the quality of online discussions by initializing (or "seeding") course materials with comments from previous iterations of the course. We hypothesize that seeding discussion forums with suitable "stimulating" posts from previous academic terms can improve students' learning gains by encouraging engagement in the forum. If true, this idea provides a practical intervention for instructors to incorporate prior to posting reading assignments, which can amplify the value and quality of student interactions in online forums. To scale this approach to large classes, we developed machine learning tools to select posts likely to stimulate discussion in future classes. Here, we show how the instructor can leverage these tools to seed online discussions. We assess the influence of this approach on students' learning outcomes and compare students' interactions with the seeds. We evaluate this process in the context of a 186-student introductory

*Kobi Gal is also affiliated with the University of Edinburgh, U.K..

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448142

biology course, using the Nota Bene (NB) collaborative annotation forum (nb.mit.edu/welcome).

Specifically, we wish to address the following questions:

(1) Can seeding change students' forum behavior, in terms of the number of participants and quality of discussion in the given thread, as well as students' engagement with course materials?
(2) Is the changed behavior linked to better learning as defined by performance on course assessments?
(3) Are some seeding methods better than others at eliciting behavior change that leads to improved learning?

Our methodology consisted of three steps: First, we designed a scale for determining the "seeding suitability" score of a given thread (an opening comment and its ensuing discussion). The scale takes into account the length and quality of the discussion, its relevance to the course material and learning objectives , as well as the level of engagement exhibited by students in the discussion.

Second, we constructed a supervised machine learning (ML) model for predicting the seeding suitability score of a given thread. The model combines lexical, emotive, and cognitive features that were extracted from students' forum interactions. We demonstrate the efficacy of the model by comparing its performance to the expert opinion of the course instructors on test/holdout data.

Third, we conducted an experiment in a live course. The ML model was used to select threads from a previous course instance. For each reading assignment in the live course, we presented a ranked list of seeding recommendations from previous courses to the live course's instructors, who could then filter the list for inconsistency and misinformation. The chosen list of seeds was subsequently seeded in the forum at the onset of the reading assignment.

We ran a randomized controlled study with three groups of students: a) students not exposed to seeds, b) students shown seeds recommended by the ML model, and c) students shown seeds recommended by an alternative model that ranked seeds purely by the length of discussion that they generated in previous course instances. For each group of students, we measured the impact of seeding on students' forum behavior, as well as on their learning gains in the course.

We found that discussion prompts from the seeding models generated more discussions than a random discussions within the Control group. Furthermore, students who received seeds selected by the ML-based model exhibited higher levels of Cognitive Engagement in their replies to seeds than those who received seeds ranked by thread length of previous discussion. In particular, we found a causal relationship between receiving seeds in the ML-based approach and improvement in learning gains.

Our results highlight the benefit of using an ML approach to augment instructors' abilities to improve students' learning when using online course forums. The ML scales up the selection process of candidate seeds in both numbers and speed, thereby providing the instructor with a manageable list of seeds for further evaluation and subsequent inclusion in the course readings.

2 RELATED WORK

The current work relates to several strands of research in learning analytics and education regarding promoting beneficial discussions in online course forums.

2.1 Interventions for Improving Learning in Online Forums

We were inspired by Miller et al. (2014), who studied the effect of seeding prior-semester comments in a physics class which contained several annotated discussion readings. They manually assessed students' comments based on a 3-point "quality" scale. Annotations lacking meaningful physics received a quality score of 0 while factual, definition type annotations received a quality score of 1. Annotations that justified questions or explanations with substantiated physics concepts received the maximum quality score of 2. As a seeding schema, for each reading, the authors randomly selected 10 first-posts that obtained a quality score of 2 as seeds. In their experiment design, different groups of students were exposed to the seeding treatment in different readings, such that all the students were exposed to the treatment by the end of the experiment. Miller et al. (2014) compared annotations from students in seeded sections versus unseeded sections by their quality measure, and by an adaptation of (Hogan et al., 1999)'s scheme developed to examine discourse patterns and collaborative scientific reasoning in peer discussions. The researchers found that students in seeded sections produce longer threads, higher quality annotations, and a greater proportion of generative threads than unseeded sections.

We extend this work by building on an ML based approach for seed selection (as opposed to the manual approach used by the authors) and by measuring the seeding impact on learning gains. This is in contrast to previous work that focused only on student discourse quality. In particular, our computational model identifies seeds that lead to perceived educational benefit.

2.2 Modeling Student Discussions

Previous studies have used a variety of models to analyze students' discussion behavior in forums and inspect the structure and quality of the discussion. In general, these works have focused on analytics rather than intervention.

Weimer and Gurevych (2007) proposed a domain-independent system for automatic quality assessment of forum posts that learns from human ratings. They constructed a computational model that uses several families of features, including surface-level (e.g., length, question or exclamation frequency), lexical (e.g., frequency of spelling errors), and similarity to the topic of the forum. They applied their approach to different web forums, highlighting situations that challenge the model, such as very short posts.

Kim et al. (2006) modeled online student discussion as a series of speech acts and investigated dependencies among the messages using a set of relational dialogue rules. They identified topics discussed in threaded discussions and assessed whether the topics shift or remain focused within the threads. They found that students who participate more and elicit more messages tend to receive better grades in the course.

Feng et al. (2006) described a system for detecting the conversation focus of threaded discussions and finding the most authoritative answer in a thread. Their model included heterogeneous evidence from different sources, such as lexical similarity, the trustworthiness of the student writing the post, and speech act analysis of human conversations. They represented discussion threads as graphs and made use of the speech act relations to generate the links. They used this model to detect which message in a thread contains the most important information (i.e., the focus of the conversation).

Several works have analyzed online forum discussions in terms of Cognitive Engagement, which is a measure to the interaction depth between the student and the course material, and has been shown to correlate with learning gains (Wang et al., 2016, Yogev et al., 2018). These works developed classification models for determining the level of Cognitive Engagement in a post. Yogev et al. (2018) showed that a visualization of Cognitive Engagement anchored in the text can give teachers valuable insight into their students' thinking and can help guide modifications of lectures and course readings to improve learning.

Geller et al. (2020) designed a computational model for detecting confusion based on rules inferred from students' hashtags (which are used to convey emotions in student posts). They showed that students' self-reported hashtags may not agree with experts' judgement about what constitutes a confused post. The authors designed computational models for automatically detecting confusion in posts that combines the perspectives of both students and experts.

All of the works above combine discourse analysis of discussion threads by expert with a supervised learning model to achieve different tasks. Although none of these works use seeding as an approach, they provide important information about the aspects of online discussions that appear to have benefited learning outcomes. These works also informed our feature design process, in particular our decision to include cognitive and emotive factors in our analysis.

3 THE NOTA BENE ANNOTATION PLATFORM

Our work builds on Nota Bene (NB), an open-source platform that lets students hold discussions in the margins of course texts, which has been used by hundreds of courses worldwide. Students highlight written passages and figures (the "marked text") and enter comments that appear to other readers in the margin. Students annotating the content may comment on or ask questions about what they are reading while classmates and instructors can reply to those comments. NB annotations are organized into threads that begin with the first-post (which can be a comment or a question), followed by all replies made by other students to the initial annotation or to the subsequent replies.

We have been studying NB's usage over 3 years in a general biology course required for all life sciences majors, many social sciences majors, and bioengineering students at a large public university in the United States. The reading materials for each lecture are created by the course instructors and are hosted on Libretexts, an online repository of open educational resources (libretexts.org). Students in the course receive reading assignments on the material uploaded to NB. The students are required to make at least three meaningful posts to the forum for each reading assignment before each class lecture. Students can satisfy this requirement by either opening a new thread or responding to another post in an existing thread. Students are also given credit for conveying their subjective opinions about the reading material by "tagging" one or more of their comments with a predefined hashtag(s) (specifically, #useful, #confused, #curious, #frustrated, #idea, #question, #help, #interested). Students receive points for this term on NB participation that count roughly 10% towards their final course grade.

Table 1: Course and forum statistics

	Winter 2018	Summer 2018	Summer 2019	Winter 2020	Summer 2020
Lectures Count	25	25	15	25	26
Num. Students	714	781	118	940	186
Num. Threads	60,469	46,315	7,278	48,228	12,746

We obtained data of students' forum interactions in four instances of the course: 1) Winter 2018, 2) Summer 2018, 3) Summer 2019 and 4) Winter 2020 terms. For all four course instances, both lecture and discussion sections were taught in-person. Data from these course instances were used to train our developed models. The Summer 2020 session of the course, in which we carried out the online seeding experiment, had 186 students and was taught online, due to COVID-19. Table 1 shows the number of students and posts collected in each course instance.

Each reading lecture in the course contains a set of instructor-defined learning objectives that the students are expected to master and the reading material contains passages relating to these learning objectives. We used pretrained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) — based on BooksCorpus (800M words) and on the English Wikipedia (2,500M words) — to learn contextual relations between students' comments and the course learning objectives. We consider a comment to relate to a given learning objective if the respective cosine similarity metric between their BERT embedding (Devlin et al., 2018) is at least 0.75. This threshold was determined in agreement with the course instructors, after observing 100 examples from the datasets of Table 1.

Table 2 shows examples of the learning objectives for reading lecture 2 of the Summer 2020 term and the numbers of posts in the lecture that relate to the learning objectives (some posts may match more than one learning objective). As shown by the table, some learning objectives are popular and dominate discussions, while others receive little attention from posts. Many posts (over a third in this lecture) do not relate to any of the learning goals, which is not captured in Table 2. If discussions were motivated to be more focused on learning objectives, this may improve students' learning in the course.

4 SELECTING THREADS FOR SEEDING

Our work is predicated on the hypothesis that it is possible to choose discussion seeds from previous semesters that can stimulate discussion and learning among current students. We therefore developed a process to help faculty *train* and then *use* a machine learning system to select good seeds from a large repository of past discussions.

Table 2: Learning objectives for a selected reading assignment in the Summer 2020 term and counts of related NB posts

Learning Objective	Num. posts matching Learning Obj.
Identify the locations of the three core atomic elements (electrons, neutrons, and protons) from a basic model model for an atom and describe their basic properties.	183
Recognize the symbols and names of the six core elements in biomolecules: C, H, N, O, P, S.	68
Interpret (identity elements and bond types) chemical and structural formulas (including 2 and 3-dimensional, of condensed, and line-angle) biomolecules.	26
Define electronegativity and explain how this concept can be used to predict the types of bonds that may be formed by two atoms.	490
Identify ionic, covalent, polar covalent bonds, hydrogen bonds, and Van der Waals interactions in different types of molecular models.	208

When reviewing students' posts, course instructors consider the following questions when deciding if a first-post may lead to a beneficial discussion for learning:

(1) To what degree does the post relate to the learning objectives of the given lecture?
(2) How much apparent effort went into writing the post, beyond the minimum requirement to get credit?
(3) Does the post raise a new idea/question that goes beyond the content presented in the reading material?
(4) How many responses did the post generate and how do the responses relate to the reading material, course objectives and others' comments?

Based on the above, we defined the *seeding score* of a thread in NB as the degree to which seeding the first post in the thread would benefit learning in the course. The seeding score is measured on an ordinal scale between 0 (not suitable for seeding) and 3 (most suitable for seeding). Two of the course instructors composed a list of general guidelines for determining the seeding score for different first-posts. An example of these guidelines are shown in the right-most column in Table 3.

The two course instructors established agreement over the labeling process by labeling 240 threads sampled from the 2018, 2019 and Winter 2020 course instances. The Kappa Interrater reliability was 0.44 (McHugh, 2012), and the instructors proceeded to resolve all of the disagreements. An additional 400 threads were subsequently labeled by one of the course instructors. The distribution of the seeding suitability score over the 640 threads is shown in Table 3. The average seeding suitability for the threads was 1.56, the median score was 1, and the standard deviation was 0.8. The low suitability reflects the need to improve the levels of discussion in the course forum. Table 4 shows a few examples of first-posts and their associated seeding suitability score, alongside the number of responses the posts generated. As shown by Table 4, there is a weak relationship between the length of the discussion and the associated seeding score for different threads.

Table 3: Distribution of seeding scores of first-posts sampled from the 2018, 2019 and Winter 2020 course instances, with examples of classification guidelines used by instructors.

Seeding Score	Amount	Percent.	Example of Classification Guidelines for Score
3	98	.153	The first-post encourages deeper understanding and critical thinking surrounding a learning objective or contains unique question/idea that brings a new perspective/discussion
2	193	.301	The first-post is related to a learning objective or the thread contain an evidence that the first-post generated productive discussion
1	320	.500	The first-post has a weak connection to a learning objective or contains a misconception/oversimplification
0	29	.045	The first-post answers a question presented in the text meant for students to answer, or detail about topics beyond the scope of class

Table 4: Examples and justification for seeding scores

First-post	Num. Responses	Seeding Score	Justification
"If energy is not made or produced, then is food the only way for our bodies to get energy? Are there any organisms that can make their own energy, if that is even possible? #question"	1	3	This question forces students to think about the inputs and outputs of the system, which is good for building mental models about energy.
"The larger the electronegativity of an atom allows the atom to attract other atoms. The most electronegative elements include Chlorine, Flourine, Oxygen, and Nitrogen."	4	1	Very basic concept without anyone in the thread even asking questions.
"My past courses focused on ecological biology, it is refreshing to finally learn molecular biology in this course. #useful"	15	0	The discussions are about course logistics.

5 COMPUTATIONAL MODELS

In this section, we describe and compare different computational models for determining seeding suitability for the first-post in a thread. The input to a model is a given reading lecture in the course, and a set of candidate threads from past course instances whose posts relate to the reading materials for the lecture. The readings in each course instance are mostly identical so we can take comments from different semesters within the same reading. The output of the model is an ordered ranking of the threads in decreasing order

of their predicted seeding score. The models used several different families of features, listed below.

Lexical and similarity based Features. This family included features relating to the counts of lexical elements in the first-post, such as the number of words in the post, presence of question mark or #question hashtag, whether it contained interrogative words (e.g., how, why) and comparative adverbs (e.g., worse, less). We also included a feature capturing the relationship between the course's learning objectives and the first-post. The feature value was the highest similarity between a first-post and any learning objective for the reading lecture. The similarity was computed as the cosine similarity between the BERT embedded representations of the learning objectives and the first-post of a candidate threads, similar to Geller et al. (2020).

Engagement Features. This family used features that build on the *Cognitive Engagement (CE)* taxonomy, which was developed by Chi (2009) to assess how deeply a student is interacting with the course material. Chi's work showed that higher cognitive engagement in students leads to learning gains. In addition, other works showed that *posts* with high CE levels are correlated with increased learning gains (Miller et al., 2016, Wang et al., 2015, Yogev et al., 2018). Thus, we posited that discussions demonstrating higher cognitive engagement might make more effective seeds that could lead to other high-engagement discussions and learning gains. We introduced some features with this goal in mind. One feature directly measured the Cognitive Engagement (CE) of posts within the candidate thread.

Posts were labeled using three levels from the CE hierarchy, in increasing levels of engagement, following the taxonomy used by Yogev et al. (2018). A post is labeled *Active* (A) if it refers to specific course materials in the annotation by paraphrasing or repeating, but does not provide new insight (e.g., "What is an isoelectric point of an amino acid?"). A post is labeled *Constructive* (C) if it displays reasoning, includes a new idea, or refers to external sources — and is not a response within a series of interactions (e.g., "The answer is c. The amino acid sequence, peptide bonds and length of the protein are derived from the primary structure...") A post is labeled *Interactive* (I) if it displays Constructive reasoning, but also builds upon, or challenges another post's ideas (e.g., "Yes, I agree with you, if the interactions between the sheets change, it may cause the protein's overall structure to change thus the function [...]").[1]

The CE level of a given post must be inferred from the text of the post and other posts in the thread. To this end, we designed a two-stage binary classification model, first distinguishing between A and C/I level posts and then distinguishing between C and I level posts. Each of the binary classification models used a Long Short Term Memory (LSTM) network structure to capture the temporal sequence exhibited by posts in the thread. The model was trained on 5,206 instances from the data sets of Table 1, tagged for CE by course experts. We evaluated the model using cross validation process that respected the temporal relationships in the data. The model achieved Precision of 0.77, Recall of 0.76, and F1 score of 0.76, which was higher than the state-of-the-art (Yogev et al., 2018) which did not use deep learning approaches.

[1]Chi's taxonomy includes a fourth, *Passive* category, but comments on the course material cannot be passive by definition.

Table 5: Example of a thread with a weighted average CE score of 0.8

Post Position	Post Text	CE Label
1st post	"Why do certain cells stay in G0 permanently? How does this affect us, since such vital cells in our body are never dividing? #question"	A
Response 1 to 1st post	"I believe this is because the cells that stay here, such as mature cardiac muscle and nerve cells, are not actively preparing to divide. There goal is not to continuously divide and make new daughter cells immediately, but rather to hold off until a signal triggers their division.#useful"	I
Response 2 to 1st post	"It makes sense that nerve cells do not divide. All the nerve cells we will ever have are usually in place by birth. This is because nerve cell development and brain development depends on the connections that nerve cells make with other nerve cells. #interested"	I
Response to Response 2	"If it were possible to turn this phase off for nerve cells and let them replicate on specific commen would it be possible to fix nerve damage in people?#question"	I

We included a feature in the seeding model that used the CE predictions to compute an average CE score for a thread. We converted a categorical scale (I, C, A) to a numeric scale by assigning points of 5, 3, and 1 respectively to each value. We computed a simple average of these points according to the thread discussion. An example of threads with a high number of I level posts that achieved a weighted average CE score of 0.8 is presented in Table 5. We note that there were extremely few A level responses in the dataset as responses generally relate to the previous comment.

Other features in the seeding model that related to CE included the number of posts in the thread exhibiting I level CE, as well as identifying that the maximum number of consecutive posts exhibiting the highest I level CE.

Another feature for describing engagement measured the likelihood of the first-post to trigger a future discussion. A straightforward approach is to use the length of the thread (the number of responses) for this purpose. Yet, we did not find a significant correlation between the thread length and the seeding score in the data we collected ($r(640) = .035, p = .187$). An alternative approach is to predict whether the first-post in the thread will incur at least one response. To this end, we trained a logistic regression classifier using BERT embedded representations of first-posts and responses from the datasets in Table 1. Evaluation of the classifier using ten-fold cross validation showed a Precision of 0.74, Recall of 0.67, and F1 score of 0.7. The correlation between the response probability obtained from this model to the seeding score value was found to be significant ($r(640) = .166, p < .001$). Examples of first-posts with different response probabilities are presented in Table 6.

Emotive Features: This family of features reflects subjective emotive information in the discussion that is conveyed by the students in form of hashtags (see section 3) and supported by the NB GUI.

Table 6: Examples of first-posts, their inferred response probabilities, and the number of responses they received

First-post Text	Response Probability	Num. Responses
"being that they have to 'couple an exergonic red/ox reaction to an energy requiring reaction in the cell' does this mean in the cell there is endergoinc processes happening? I'm just confused about this sentence. does anyone have an example?"	0.632	1
"this is similar to the interactions between the amino acids that make up the proteins. hydrogen bonds are clearly very important in the structural stability of not only proteins but also nucleic acids. just one way the molecular structure or the chemisty is able to describe biological occurences. #idea"	0.358	2
"I appreciate how this brings up all products and results of the reaction, rather than just what is deemed "relevenat" for the course. I appreciate knowing the detials of the net loss and the precursors!"	0.294	0

We hypothesizd that expressing emotions may invite other students to participate in the discussion and can indicate good threads for learning.

We explored the correlation between different hashtags occurrences count in a thread discussion and the target variable (seeding score) and found that only specific hashtags has a significant correlation: #curious and #frustrated. Therefore we focused on those hashtags alone as well as whether the post expresses confusion. We note that the educational literature shows that students' posts may reflect confusion without self-reporting this hashtag, for example, when they are stating something incorrect that they seem to think is correct (Plaut, 2006). To this end we used the classifier by Geller et al. (2020), which extracted features from the post and the selected text in the reading material to infer instances of student confusion that go beyond the use of #confused hasthtag.

Table 7 summarizes the features within the different families, including the correlation between the features and the seeding score that was determined by the experts on the 640 labeled threads. The most important features with positive contribution to seeding score were whether the first-post was a question, the response probability, and whether or not the longest sequential CE level in the thread discussion is interactive (as in Table 5). The most important features with negative contributions to seeding were whether the first-post expresses an incorrect statement without asking a question, the percentage of posts in the thread with fewer than 4 words, and the number of #frustrated hashtags in the thread discussion.

5.1 Offline Evaluation

In this section, we compare the performance of different computational models for predicting whether a post is worthy of seeding. We trained several computational models on the data, using the features described above and the labels assigned by instructors as described earlier. We present here the two models that outperformed all the others. The first computational model is a binary classifier using Linear Discriminant Analysis. The second is a regression model using Ridge regression. Both models used the features described in the previous section, and received a candidate thread for seeding as input. The binary classifier returned the probability of the first-post in the thread being suitable for seeding, according to the criteria described below, while the regression model predicted the seeding score of the first-post in the thread. The regression was found to be significant ($r(640) = .306, p < .004$), with the following features receiving the highest coefficients values: whether the first-post contains a question, the number of #curious hashtags in the thread discussion, the probability of the first-post triggering a response, and whether the first-post exhibited a Constructive CE label.

We evaluated the two approaches over all thread instances using ten-fold cross validation. The number of training instances in each fold was 576 and the number of test instances was 64. The gold standard was whether a given thread is suitable for seeding. In accordance with the course instructors, we determined that threads with a seeding score of 2 or 3 — accounting for 45 percent of all threads — were suitable for seeding. For the regression model, the average seeding score on the training set was 1.56 (equal to the average score provided by the course instructors), the median score was 1, and the standard deviation was 0.8. In this model, a seeding score of above 1.53 accounted for 45% of all threads. Hence, the regression model used the 1.53 value as the lower bound to determine whether a thread was relevant for seeding. The binary classifier used the 2.0 value as the lower bound, as determined by the course experts.

Table 8 compares the two models using traditional measures of information retrieval, including F1 Score, Accuracy, and AUC. As shown in the table, the binary classifier provided a higher score in all measures. We note that in the context of seeding of suitable threads, false positives (initializing the forum with a bad seed) are more critical than false negatives (leaving out a suitable seed). Thus, precision is the most important metric.

Given the above results, one would consider preferring the binary classifier approach. Nonetheless, for the seeding problem we need to consider a ranking based approach which reflects our intended methodology to provide recommendations to course staff about which threads to seed for each reading lecture. We evaluated how the models compare with the instructors in terms of alignment with the top-n threads that was determined by the instructors. For each fold, we ordered the threads in the test set according to predicted seeding score (for the regression model) and according to predicted probability for suitability (for the classifier model). Table 9 compares the alignment of the top n-scoring threads in each model based on the instructors true labels for $n = 15$ and $n = 25$, which reflected the number of possible seeding recommendations to provide the course instructors. As shown by the table, the regression model outperformed the binary classifier in all these measures; this includes the Normalized Discounted Cumulative Gain (NDCG), which penalizes highly relevant instances appearing lower in the ranking list, and is commonly used to evaluate recommendation systems.

Given the superiority of the regression method in the ranking task, we decided to use it for our online study, which envisions to provide a ranked list of seeding recommendations to instructors.

Table 7: Pearson's Correlation between Seeding score and suggested features

Feature Family	Feature Description	Corr.	p-value
Lexical	Question identifier in the first-post	0.192	0
Engagement	First-post response probability	0.166	0
Engagement	The longest sequential cognitive engagement level in the thread discussion is I level	0.135	0
Emotive	Counts of #curious hashtag in the thread discussion	0.133	0
Lexical	Counts of Wh-adverb POS tag in the first-post such as: how, when, where, why	0.128	0.01
Emotive	The first-post express confusion and contain a question	0.126	0.01
Engagement	Thread discussion cognitive engagement score	0.122	0.001
Lexical	Average number of words used in thread discussion	0.094	0.008
Lexical	First-post similarity to a learning objective	0.089	0.02
Engagement	Counts of posts with I cognitive engagement level in thread discussion	0.085	0.016
Lexical	Counts of Adverb comparative POS tag in the first-post such as: worse, less, better	0.079	0.02
Lexical	Counts of particles POS tag in the first-post such as: not, 's	0.074	0.03
Lexical	Number of words used in first-post	0.063	0.08
Engagement	First-post labeled with C cognitive engagement level	0.063	0.055
Lexical	Counts of Wh-determiner POS tag in the first-post such as: what, which	0.05	0.1
Emotive	The first-post express confusion and contain a statement	-0.120	0.001
Lexical	The percent of short posts (less than 4 words) in thread discussion	-0.108	.003
Emotive	Counts of #frustrated hashtag in the thread discussion	-0.103	0.005
Emotive	The percent of posts that contain only hashtags in thread discussion	-0.069	0.04

Table 8: Classification results comparison

Seeding Model	Precision	Recall	F1 Score	Accuracy	AUC
Binary classifier	**0.67**	**0.64**	**0.64**	**0.64**	**0.63**
Regression model	0.66	0.61	0.62	0.61	0.62

Table 9: Ranking results comparison

Seeding Model	Precision @15	F1 score @15	Precision @25	F1 score @25	NDCG
Binary classifier	0.64	0.45	0.60	0.56	0.86
Regression model	**0.66**	**0.46**	**0.61**	**0.57**	**0.89**

6 ONLINE SEEDING EXPERIMENT

In this section, we apply the computational model described in the previous section in an active course. We focused on the summer 2020 instance of Bis2A, which was conducted online due to COVID-19. The course included 26 reading lectures, two midterm exams (after reading lecture 10 and after reading lecture 18) and a final (after reading lecture 26).

Students enrolled into a unique recitation/discussion section (20-24 students), each of which is moderated by a graduate student teaching assistant (TA). Each section used its own instance of NB and did not see the comments of the other sections. In the first half of the course (readings 1-12) we did not publish any seeds. The first midterm (hence referred to as Midterm 1) following reading lecture 10 was used as a baseline for comparing course performance between the different groups. Seeding commenced in the second half of the course (from reading lecture 13 to reading lecture 26). This permitted a *difference in differences analysis*, in which we compared the impact of a *change* at midpoint from the control condition to seeding.

The study was reviewed by the IRB of the hosting institution and deemed exempt. All students who participated in the online study in summer 2020 filled a consent form and were requested to opt-in the study.

6.1 Experimental methodology

Our seeding methodology consisted of four steps: (1) For each reading lecture in which seeding is performed, select a set of candidate seeds from the first-posts of the relevant lecture in past instances of the course. All of the candidate seeds were selected from past instances of the course in 2018, 2019, and Winter 2020 from Table 1. (2) Rank the candidate seeds according to a scoring function. (3) Present a ranked list of the top scoring 15 seeds to the course instructors and allow the course instructors to review the ranked list to select a subset of ten seeds for publishing. (4) Publish the seeds in the course forum in the relevant reading lecture. We expand on each of these steps in turn.

Step 1: The instructors selected a subset of learning objectives that fulfilled the following criteria: Each learning objective was related to readings in the first and second half of the course; as

well as to questions in midterm 1 and the final exam. For each of these learning objectives, we selected the 50 first-posts with the highest similarity to the learning objective, computed using BERT embedding, and added these posts to the candidate seed set. As a preprocessing step, we removed threads whose first-post met one or more of the following: 1) selected text in the readings that was not in the given lecture reading, 2) did not contain any text, or 3) only contained a hashtag. In practice, there were between 200 and 250 candidates for each lecture.

Step 2: We used two different approaches to rank the candidate set of seeds from the previous step. The *ML-supported* approach used the regression model described in the previous section to rank candidate seeds according to their predicted seeding score. The *Long* approach ranked seeds according to the length of the discussion that they generated, in terms of number of responses. We selected this baseline approach following previous studies that measured thread quality by the number of responses (Kim et al., 2006). In addition, The Long approach does not require any AI, and is simple to compute. Both approaches broke ties by ordering with respect to the semantic similarity between the seed and the learning objectives of the reading lecture. To ensure that the selected seeds for the lecture are all consistent with a single poster, both approaches removed seeds whose marked text in the reading material overlaps with a higher ranking seed.

Step 3: The course instructors inspected the ranked list of 15 seeds for each seeding group using their expertise. For example, instructors identified seeds with inconsistencies or contradictions with another seed, in which case the lower ranked seed was removed. Instructors also removed seeds that explicitly referenced to previous course instances. Grammar correction was purposely not done. In practice, 38 of 180 seeds were rejected by the instructor in the ML condition and 34 of 180 seeds were rejected in the Long condition. The ten highest scoring seeds that passed the instructors' inspection were subsequently planted in the NB forum. An example of contradicting seeds is presented in Table 10.

Step 4: The chosen ten seeds were manually inserted in NB to coincide with the time that the reading materials for the following lecture was published. Informing the students that the seeds were selected by the course staff could bias their response. There is evidence showing that students respond more positively to questions posed by fellow students (Mazzolini and Maddison, 2003). Therefore, all published seeds were tagged with a fictitious student account assigned with the gender, race, and ethnically ambiguous name "Lee Jordan". The distribution of the location of the seeds in the reading material for the lecture was similar in both seeded conditions, in that most of the threads were positioned in the first 50% of the lecture. This trend was also exhibited by students in past course instances.

Our methodology was evaluated using a randomized control study. Discussion sections in the course were assigned to different experimental conditions. Students in the ML-seeded condition (59 students) and the Long-seeded condition (59 students) were assigned seeds ranked by the ML-supported and Long approaches, respectively. Students in the Control condition (68 students) were not exposed to seeds.

Table 10: Example of contradicting seeds

Seed Text	Predicted Seeding Score
"It is clear how the membrane comes to be charged, but not what purpose the polarized membrane serves. does the polarization help to move the electrons along the membrane? this however wouldn't make sense if it is the electrons that cause this polarization. #confused"	2.15 (accepted)
"Basically, the proton motive force works as a tiny battery. it's energy can either be stored for later use in atp or can be used immediately to do work like powering the movement of flagella . this proton motive force occurs when the embedded electron carriers in the cell membrane become energized as a result of the electron transport reactions."	1.65 (rejected)

6.2 Hypothesis and Results

In this section we evaluate the effect of seeding on students' learning gains in the course, Before stating our main hypothesis we make the following definitions. First, we use the term "seeded learning objectives" to refer to those learning objectives that were selected by the course staff for seeding in both of ML- and Long-seeding conditions (See Step 1 in Section 6.1). All of the seeds planted in the study related to at least one of the seeded learning objectives. Second, we define the set of "relevant" questions in an exam as those questions that related to the seeded learning objectives. There were 7 relevant questions in midterm 1 and 26 relevant questions in the final exam. We can now state our hypothesis which relates to the difference in the learning gains obtained by students in the ML-seeded condition and the control condition.

Learning gains hypothesis: Students in the ML-seeded condition will incur higher learning gains than students in the Control condition, as measured by their performance in relevant questions in the final exam. (The null hypothesis states that the learning gains in the ML-seeded condition will not be higher than in the control.)

To address this hypothesis, we first compared the success rate (the ratio of correct answers) on relevant exam questions between students in both ML- and Long-seeding conditions. We did not find a significant difference between the ML-seeded condition and the control in midterm 1 which occurred prior to the beginning of treatment (Repeated Measure ANOVA test ($F(2, 12) = 2.240, p = 0.149$)). This implies there was no difference in students' knowledge between the different conditions prior to beginning the seeding intervention.

We found a significant difference between the condition groups (ML, Long, and Control) in the final exam (Repeated Measure ANOVA test ($F(2, 50) = 3.818, p < 0.03$)) after the treatment. A post hoc test revealed that the average success rate for the ML-seeded condition (0.69, STD= 0.21) was higher than that of the Control condition (0.64, STD=0.20) ($p < .03$). Thus, we can reject the null hypothesis in favor of the learning gains hypothesis. We do note that the average success rate of the ML-seeded condition was not significantly higher than that of the Long-seeded condition (0.66, STD=0.21) ($p = .09$).

We also wished to determine whether there exists a causal effect of seeding for both seeding conditions. To this end, we used the difference-in-differences approach (Lechner et al., 2011). The independent variables for the model were as follows: Treatment (whether seeding or control), Time (whether midterm 1 or final exam), Treatment*Time (1 when Treatment is seeding and Time is the final exam; 0 otherwise). The dependent variable was the students' success rate in the relevant questions in midterm 1 or the final exam, depending on the value of Time. We controlled for the effect of a confounding variable (the identity of the Teaching Assistant) by clustering the standard errors (Bertrand et al., 2004, Wooldridge, 2003).

The difference-in-differences analysis between all the treatment groups (ML-seeded and Long-seeded) to the Control group revealed that both seeding treatments had a significant effect on students' learning gains, with a coefficient value of 5.517 for the Treatment*Time ($p < 0.001$). While focusing on the ML-seeded group, The difference-in-differences analysis between the ML-seeded and Control conditions showed a coefficient value of 5.28 for the Treatment*Time ($p = 0.07$), which implies that a causal relationship is possible between the ML-recommended seeds and students' learning gains. However, when focusing on the Long-seeded group the difference-in-differences analysis between the Long-seeded and Control conditions we did not find a meaningful effect on learning gains.

We also provide the following results of an exploratory analysis of the seeding effects on several aspects of forum behavior of students in the different conditions.

Amount of discussion: We found that the number of responses to seeds in the ML condition (177 responses) was 16% higher than the number of responses to the seeds in the Long condition (153 responses, proportion test $z = 1.795, p = 0.073$) and 22% higher than the number of responses in the Control threads (145 responses, proportion test $z = 2.398, p < 0.016$). (The results in the Control condition are based on a set of 10 randomly sampled threads for each lecture, the same number of seeds that were planted in each of the seeding conditions).

Length of discussions: The average length of responses (measured in the number of words) to seeds in the ML-seeded condition (52.75, STD= 26.10) was found to be higher than that of the Long-seeded condition (43.02, STD= 25.82) and the Control condition (43.02, STD= 22.27). There was a significant difference between the condition groups (Kruskal-Wallis test ($X2(2) = 15.588, p < 0.001$). Post hoc analysis showed that the average response length in the ML-seeded condition is higher than that of Long-seeded ($p < 0.001$) and Control ($p < 0.024$), and that the average word length was higher in the Control condition than in the Long-seeded condition ($p < 0.025$).

Quality of discussion: The level of Cognitive Engagement of posts made in reply to seeds was found to be highest in the ML-seeded condition, in comparison to the other two condition groups. We found that the number of posts that exhibited the highest CE level (interactive) was higher in the ML-seeded condition (112) than in the Long-seeded condition (54) (Chi-Square ($X2(1) = 24.597, p < 0.001$) and in the Control condition (49) (Chi-Square ($X2(1) = 26.547, p < 0.001$). In the end, ML seeds generated twice as many interactive responses compared to the baseline approach and compared to randomly sampled threads in the forum.

6.3 Discussion and Conclusion

The goal of this study was to develop automatic support for course instructors who wish to engage students in meaningful online discussions on course material. Using an AI approach to generate candidate seeds from prior students allows for the scaling up of the selection process of candidate seeds in both numbers and speed and provides a ranked list of recommended seeds for the instructor. Rather than attempting to be fully automatic, the approach described in this study relies heavily on the involvement of human experts (e.g., the course instructors) for all steps. Instructors are engaged in designing a measure for scoring seeds according to educational impact, using the measure to label candidate posts for seeding, inspecting the proposed list from the algorithm, and selecting the chosen set of candidate seeds to plant in the forum.

In this respect, our approach occupies a "sweet spot" in the continuum between providing no support and fully automated support for seeding, in which the algorithm and the human expert share the problem solving, each performing tasks for which they are best suited. Here we are demonstrating a prototype for Human-AI collaboration that can be experimented with other forms of task allocation, such as allowing the instructor to override the machine generated recommendations with their own choices, or to use active learning methods to improve the selection algorithm over time.

Our results suggest that seeding can have a positive impact on commenting behavior by students, that the method used for selecting seeds makes a difference, and that some of these changes may even be associated with the potential for learning gains. Both ML-seeded and Long-seeded approaches changed students' forum behavior in ways that led to more posts, longer threads, higher CE, and better learning gains.

Also of note is the apparent advantage of the ML-seeded approach over the Long-seeded approach with regard to the quality of the ensuing discussion and students' learning gains. We found that students in the ML-seeded condition exhibited greater cognitive engagement in their posts than did students in the Long-seeded condition. We also found evidence for a possible causal association (using the difference-in-difference analysis) between the ML-seeded approach and student learning gains, when constraining our analysis to exam questions related to the specific learning objectives for which we planted seeds. This means that it is worthwhile to invest effort in understanding the reasoning and criteria used by instructors to choose good seeds, and to use that knowledge to inform the design of automatic models for supporting their work.

We did not find a difference in performance between students who actively responded to seeds and other students in the same group. This implies that part of the effect may be on account of students reading, but not participating in the discussion in the forum. This is encouraging, as students are significantly more likely to read comments than they are to write comments.

In the end, the ML classifiers developed in this study were able to take a large comment corpus and deliver a subset of suitable candidate seeds using lexical, engagement and emotive based features

of posts (Table 5), properties which are both informative and generalizable to other contexts and disciplines. Aside from re-posting comments as seeds, these comments may serve valuable purposes in instructional settings, for instance, as discussion questions or as questions to be embedded within the course text itself.

Finally, we mention several limitations of our work. First, although the benefits of the seeding approaches for student engagement and performance are clear cut, the increase in number of responses was not large. This can be traced back to the way NB is used in the course. Students generally choose to commence new threads (e.g., asking a question about or paraphrasing the material) rather than responding to existing threads, which requires more effort. Therefore, the signal from the intervention study is not strong. It may be that providing more feedback from course instructors, or requiring students to reply to other threads, will motivate students to be more responsive to others.

Second, our work was done using the NB platform, which differs from many forums in that the discussions happen in the margins of the course material rather than a separate forum. We can offer no evidence that our approach will be effective in these more traditional platforms, although we believe that it would.

For future work, we wish to use our model outputs (recommended seeds) to promote discussions in other contexts, such as for discussion topics in small groups, during class, to be put into the text itself, etc. Also, we wish to employ an active learning approach in which feedback from the instructors regarding good and bad seeds is used to update the model after each lecture.

REFERENCES

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* 119, 1 (2004), 249–275.

Lori Breslow, David E Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and Daniel T Seaton. 2013. Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment* 8 (2013), 13–25.

MTH Chi. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1 (1), 73–105.

Dave Cormier and George Siemens. 2010. The open course: Through the open door–open courses as research, learning, and engagement. *EDUCAUSE review* 45, 4 (2010).

Gerardine DeSanctis, Anne-Laure Fayard, Michael Roach, and Lu Jiang. 2003. Learning in online forums. *European Management Journal* 21, 5 (2003), 565–577.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. 208–215.

Shay A Geller, Nicholas Hoernle, Kobi Gal, Avi Segal, Amy X Zhang, David Karger, Marc T Facciotti, and Michele Igo. 2020. # Confused and beyond: detecting confusion in course forums using students' hashtags. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 589–594.

Kathleen Hogan, Bonnie K Nastasi, and Michael Pressley. 1999. Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and instruction* 17, 4 (1999), 379–432.

Jihie Kim, Erin Shaw, Donghui Feng, Carole Beal, and Eduard Hovy. 2006. Modeling and assessing student activities in on-line discussions. In *Proc. of the AAAI Workshop on Educational Data Mining*. 16–17.

Michael Lechner et al. 2011. *The estimation of causal effects by difference-in-difference methods*. Now.

V Light, E Nesbitt, P Light, and JR Burns. 2000. 'Let's You and Me Have a Little Discussion': Computer mediated communication in support of campus-based university courses. *Studies in Higher Education* 25, 1 (2000), 85–96.

Margaret Mazzolini and Sarah Maddison. 2003. Sage, guide or ghost? The effect of instructor intervention on student participation in online discussion forums. *Computers & Education* 40, 3 (2003), 237–253.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

Kelly Miller, Sacha Zyto, David Karger, and Eric Mazur. 2014. Improving online class forums by seeding discussions and managing section size. In *Proceedings of the first ACM conference on Learning@ scale conference*. 173–174.

Kelly Miller, Sacha Zyto, David Karger, Junehee Yoo, and Eric Mazur. 2016. Analysis of student engagement in an online annotation system in the context of a flipped introductory physics class. *Phys. Rev. Phys. Educ. Res.* 12 (Dec 2016), 020143. Issue 2. https://doi.org/10.1103/PhysRevPhysEducRes.12.020143

Suzanne Plaut. 2006. "I Just Don't Get It": Teachers' and Students' Conceptions of Confusion and Implications for Teaching and Learning in the High School English Classroom. *Curriculum Inquiry* 36, 4 (2006), 391–421.

Dawn M Poole. 2000. Student participation in a discussion-oriented online course: A case study. *Journal of research on computing in education* 33, 2 (2000), 162–177.

L Romeo. 2001. Asynchronous environment for teaching and learning: Literacy trends and issues online. *Delta Kappa Gamma Bulletin* 67, 3 (2001), 24–28.

Xu Wang, Miaomiao Wen, and Carolyn P Rosé. 2016. Towards triggering higher-order thinking behaviors in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 398–407.

Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé. 2015. Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society* (2015).

Markus Weimer and Iryna Gurevych. 2007. Predicting the perceived quality of web forum posts. In *Proceedings of the conference on recent advances in natural language processing*. 643–648.

Jeffrey M Wooldridge. 2003. Cluster-sample methods in applied econometrics. *American Economic Review* 93, 2 (2003), 133–138.

Eran Yogev, Kobi Gal, David Karger, Marc T Facciotti, and Michele Igo. 2018. Classifying and visualizing students' cognitive engagement in course readings. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, 52.