Classifying and Visualizing Students' Cognitive Engagement in Course Readings

Eran Yogev and Kobi Gal
Ben-Gurion University
kobig@bgu.ac.il and
yogeveran@gmail.com

David Karger
MIT CSAIL
karger@csail.mit.edu

Marc T. Facciotti and
Michele Igo
UC Davis
mtfacciotti@ucdavis.edu and
mmigo@ucdavis.edu

ABSTRACT

Reading material has been part of course teaching for centuries, but until recently students' engagement with that reading, and its effect on their learning, has been difficult for teachers to assess. In this article, we explore the idea of examining *cognitive engagement*—a measure of how deeply a student is thinking about course material, which has been shown to correlate with learning gains—as it *varies* over different sections of the course reading material. We show that a combination of *automatic classification* and *visualization* of cognitive engagement anchored in the text can give *teachers*—and not only researchers—valuable insight into their students' thinking, suggesting ways to modify their lectures and their course readings to improve learning. We demonstrate this approach with analyzing students' comments in two different courses (Physics and Biology) using the Nota Bene annotation platform.

INTRODUCTION

Reading material has long been an integral part of many courses. Primary sources and carefully organized textbooks can enrich and complement the limited amount of material that can be presented during lectures. But unlike the (ideally) two-way interactions of lectures, where a teacher can get some sense of how well the students are engaged with and understanding the material, reading happens out of sight, leaving the teacher with relatively little sense of how or even whether the reading is helping students to learn.

Indeed, there is evidence that passively reading is a relatively poor way to learn. Chi and Wiley [7] developed the *ICAP* framework to taxonomize different levels of learning engagement for various classroom activities. This framework orders learning into *passive* engagement like reading a text (or listening to a lecture), *active* engagement (taking notes), *constructive* engagement (reasoning and solving problems), and *interactive* engagement (discussing the material with others). Through controlled experiments, Chi and Wylie demonstrated

that activities higher in the cognitive engagement ordering led to better learning gains.

This paper studies how cognitive engagement (CE) varies over different sections of the course reading material, and whether it informs *teachers*, and not only researchers, about the efficacy of their teaching. We argue that a combination of *automatic classification* and *visualization* of CE anchored in the reading text can give teachers valuable insight into their students' thinking, suggesting ways to modify their lectures and their course readings to improve learning.

Our approach leverages an *annotation-based* online discussion forum, where students comments about the reading are placed in the margins of (specific parts of) the course text. Working with teachers of large biology and physics courses, we assess the levels of CE demonstrated by these comments and examine how CE varies over the text. In two separate courses, we show that using features of the annotated text can significantly improve the automatic classification of CE over prior work that only used the content of the comments themselves. We develop a heat map visualization that presents variations in CE over the text, and demonstrate its efficacy for teachers through a user study.

Online Forums and CE

Online discussion forums offer the promise of a new way to increase the highest, interactive mode of engagement with course material, and have been shown to improve learning gains [4, 8, 9]. In both online courses [19, 14] and traditional ones [13], online discussion forums help teachers track students' learning throughout the course, to identify points of difficulty or confusion, or to actively guide the student discussion in a pedagogically useful way.

While online discussion forums can be powerful pedagogical tools for students and instructors alike, the volume of content generated from frequent (multiple times per week) use in large-enrollment (>200 students) courses challenges the instructional staff to analyze, learn from, and respond productively to student comments in a timely manner. For example, students in the first-year introductory biology course (BIS2A) offered at UC Davis, generated in excess of 3000 comments every other day (>9000 per week) using an online discussion tool [22]. Computational tools that help streamline the instructional staff's analysis of the online discussion are therefore highly desirable. To this end, Wang et al. [16, 17] extended the use of the ICAP framework to analyze discussion forums

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

L@S 2018, June 26–28, 2018, London, United Kingdom

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-5886-6/18/06...\$15.00

DOI: <https://doi.org/10.1145/3231644.3231648>

in a MOOC setting. They showed that high levels of CE, as reflected in students' behavior in the forum, are positively correlated with students' learning gains.

Our Work

This paper describes our work to develop and evaluate new computational tools focused on detecting, from their forum discussion behavior, the degree to which students engage cognitively and meaningfully with the course materials. The resulting tools enable teachers to distinguish students who are actively involved in analyzing and synthesizing course information from those who are interacting with the information more passively. The information provided by the in-place commenting forum can also help instructors to distinguish areas within the text that trigger deeper student engagement from areas that may need editing for clarity or enrichment. Finally, information on CE derived from forum posts can aid teachers' class planning and can also inform their decision of whether and where to actively participate in the forum to facilitate student learning.

We focus our investigation on an in-place discussion forum called *Nota Bene* (NB) in which students can anchor their comments within the course material during their learning. Prior work [22, 13] has shown that the context provided by such forums allows students to consider all of the relevant discussion threads, improves the quality of the discussion and promotes student-to-student feedback (when compared to traditional temporal based forums).

Our contribution is threefold: First, we analyze the CE anchored in text. For two courses each involving hundreds of students, teaching staff examined and labeled CE for a corpus of several thousand in-place comments. Manually labeling students' annotations for CE is effort-intensive, requiring time and expertise in pedagogical practices and the reading material [16]. We developed an *interactive decision tree* (derived from Wang et al. [16]) to speed up and simplify this labeling.

We examined the labeled data to draw conclusions about how anchoring comments in text changes CE. For example, we found that it dramatically reduced the amount of (lowest level) passive CE because most comments were forced to engage at least to some degree with the text they were annotating. This differs from traditional forums where such off-topic behavior is common [14, 10]. Anchoring in text also provides valuable information about students' engagement with the material that is not possible in traditional forums. For example, CE can identify cases where students' comments paraphrase the text book rather providing their own insight. We also find that different sections of the text seem to encourage different levels of CE—for example, that definitions in the text generate relatively low CE.

Second, we provide machine-learning classifier of CE in annotation based forums. To do so we extend the bag-of-words model used by Wang et al. to consider additional features linking the students' comments to the content they are annotating. We show that the context provided by the in-place discussion forum significantly improves the performance of the CE classifier in two separate courses (Biology and Physics).

Third, we show that CE can serve as a useful tool for teachers aiming to improve their teaching, and not just researchers. We developed a visualization tool which displays how the CE of comments is distributed over the course textbook/course materials. The tool provides a heat-map that allows course staff to identify areas of importance in the course textbook (e.g., with extreme high or low CE value) and “zoom in” on these areas to uncover how CE scores relate to specific student comments. We conducted an extensive user study in which the course staff of the biology course evaluated the heat map on real data and more generally, whether CE could be used as a tool to benefit their teaching practices both in and outside the classroom. All course staff agreed that visualizing CE using a heatmap could aid their teaching. Most of the staff would use it to identify students' annotations exhibiting high or low CE values, which they would use to facilitate their decision of whether to intervene in the forum. We also discovered that the course instructors and the teaching assistants (TAs) would use the CE heatmaps in different ways. Course instructors stated they would focus on discussion threads exhibiting high CE that reflect heightened interest by students, and the instructors stated that they would use this information to help plan the class. TAs stated that they would look for areas exhibiting extreme values of both high and low cognitive engagement, so that they can relate to individual students during section.

Our results provide valuable insights to course designers, suggesting that analyzing students' forum based behavior using CE can aid teaching, and that using in-place annotation forums offer a natural way to bring out these benefits.

RELATED WORK

Chi and Wylie [7] defined the notion of CE and hypothesized that the degree of student learning depends on the level of engagement in the CE hierarchy, which is conveyed in classroom behavior. They verified this hypothesis using the ICAP framework in a lab and classroom study, in which students that were assigned with behaviors that were more engaged with the learning materials also exhibited higher learning gains.

Zue [21] studied students' CE behavior in four different social science classes. They show that the instructor's presence in the forum, a student's prior knowledge, motivation, and interest in the topic influences their cognitive engagement as it is reflected in their discussion behavior. In particular, students with more prior knowledge of the material are more bored than students with less, causing them to write less meaningful comments.

Several works studied CE using computational models. Wang et al. [17] explored the relationship between students' CE, as reflected in the messages they write in course forums, and their learning gains. They found that on-topic discussion (which includes posts about readings, quizzes and assignments) was associated with higher learning gains than off-task discussion (which includes posts about administrative issues in the course and social networking). They provided a bag-of-words classification model for CE that we extend to in-place annotation forums in this work.

In subsequent work, Wang et al. [16] adapted the definition of Chi and Wylie [7] to a MOOC forum. They created a manual

Figure 1. The Nota Bene interface. Left: Course document; Right: Annotations

for labeling user messages in the forum to their respective cognitive engagement category. For each such category, the manual supplied a definition, situations where this definition applies and multiple examples for correct and incorrect labeling. They employed several measures, including regression analysis, propensity score matching [5], and a within-subject analysis to study whether higher order CE behavior leads to better learning in the course. They showed that students who contributed at least one comment exhibiting high CE exhibited higher learning gains as measured by than those that did not. They also studied which course topics from the textbook trigger richer discussion using Latent Dirichlet Allocation (LDA). They showed that topics containing words from people’s daily experience are more positively associated with a richer discussion than technical terms that people do not frequently use in daily life.

Lastly, Wen et al. [18] measured learner motivation using cognitive engagement and word abstractness [15] as features. They used these measures in survival models in three MOOCs to predict if students are about to leave the course. They found that students whose posts expressed an average of one standard deviation more learner motivation are 42 percent more likely to continue posting in the course forum, and those whose posts have an average of one standard deviation higher CE level are 6 percent more likely to continue posting in the course forum.

THE NOTA BENE ANNOTATION TOOL

Our empirical methodology is based on the Nota Bene (NB) [22] platform, an open-source collaborative online annotation tool for PDF, HTML, and video files. The course content (e.g., textbook) is uploaded to the NB website, to which students log on to read and annotate the text. Annotations are made by highlighting a passage of the textbook (called “the marked text”) and typing into a text field that appears in the margin. Students annotating the content may comment or ask questions about what they are reading; classmates can reply to those comments. NB Annotations are organized into *threads* which constitute a starting comment or question followed by all the replies made by other students to the initial annotation or to the subsequent replies.

Figure 1 shows the NB interface for a Biology course that is part of our empirical methodology. On the left is a snippet from the course textbook about Amino Acids and Protein Structure. The reading material is augmented by annotations that the students and faculty have written, which appear as expandable discussions on the right-hand-side panel. Annotations are anchored to particular locations in the document. Clicking on the document (for example on “Amino Acids”, appearing in blue in the figure) scrolls to the corresponding annotation. The reading on the left embeds a heat map which highlights areas in the document associated with many (dark yellow) and few (light yellow) annotations.

The in-place structure of the NB tool allows students to interact in the forum while they are reading the course material and provides context to the discussion. Use of NB has been shown to positively contribute to students’ contributions to the forum, promote student-to-student feedback, and encourage learning [13].

For the purpose of this study we used NB in the context of two courses. The first course was the BIS2A course for introductory biology taught at UC Davis. The course included ≈ 1000 students, one instructor, one course coordinator, and ≈ 17 teaching assistants every quarter. The instructor lectures to students in sections of ≈ 500 while TAs see students in sections of ≈ 24 . The second course was an introductory, calculus-based physics course called AP50, intended for engineering and premedical students taught at Harvard University. AP50 is split into two courses taught consecutively, AP50A (mechanics), including 77 students, and AP50B (electricity and magnetism), including 73 students. The course was taught using a flipped classroom approach. For both biology and physics courses, students receive reading assignments on material which is uploaded to NB. The students are required to provide a meaningful comment to the reading assignment in NB before each class and to actively participate in discussions.

ANALYZING COGNITIVE ENGAGEMENT

We use the extended taxonomy of Cognitive Engagement (CE) by Wang et al. [16] for assessing online discussion. Table 1 shows the extended CE taxonomy table, listing the different categories on an increasing ordinal scale.

We illustrate the categories using examples from the biology course we study in this paper. Annotations exhibiting Off-task (O) behavior do not refer to the reading directly nor to student’s experience with the reading (e.g., “The picture in page 6 makes me think of a motorboat and its wake[...]).

Annotation exhibiting Active-general (A2) behavior indicates attention to the course in some way without referring to content (e.g., “Actually, I think the videos are sometimes more useful than the reading, because they have more examples, and because it is a video[...]).

Annotations exhibiting Active-targeted (A1) behavior displays engagement with specific course materials in the annotation by paraphrasing, repeating, or mapping resources, but does not provide insight (e.g., “What is an isoelectric point of an amino acid?”).

Table 1. Extended CE Taxonomy (Wang et al.)

Label	Description
O	Off Topic
A2	Active General
A1	Active Targeted
C2	Constructive Extending
C1	Constructive Reasoning
I	Interactive

Annotations exhibiting Constructive-Extending (C2) behavior introduce a statement or question related to the reading that includes a new idea or refers to external sources, but does not reason about them (e.g., “Even though these three parts of protein folding are classified separately, it does not necessarily mean that they occur independently.”).

Annotations exhibiting Constructive-Reasoning (C1) behavior display reasoning about the reading, for example by explaining a phenomenon, stating a cause and effect, or drawing a conclusion (e.g., based on scientific evidence or personal experience), but don’t interact with previous posts (e.g., “The answer is c. The amino acid sequence, peptide bonds and length of the protein are derived from the primary structure, which is maintained in denaturation[...].”).

Annotations exhibiting Interactive (I) behavior displays Constructive Reasoning (C1), but also responds to a previous annotation, for example, by elaborating on, pointing to, building upon, or challenging the other 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. The change in the sheets also can change the overall polarity of the molecule, thus can change how it will react with water [...]”).

Previous work on CE included domains such as classrooms and forums [7, 17, 16, 18]. In all of these works, determining CE was based solely on students’ comments in the forum. However, annotations in Nota Bene are anchored to selections in the textbook. This can provide context that can help to infer the cognitive state that is embedded in the comment. For example, by comparing the selected text with the relevant annotation, we may be able to distinguish a comment that is adding meaningful content from one that is only paraphrasing the original course content.

We adapted the manual by Wang et al. to coding CE for annotated based forums. We extended the definitions for the CE taxonomy to incorporate the marked text and its surrounding paragraph in addition to the annotation itself. The most important modifications were as follows: (1) If the annotation is a conclusion that derives from the marked text or the surrounding paragraph, then the CE classification of the annotation is Constructive-reasoning; (2) If an annotation is paraphrasing the marked text/paragraph, then the CE classification of the annotation is considered Active-targeted. The maximal depth of the tree (and also the number of questions) is six (the same as the depth of the decision-tree used by Wang et al. [16].)

For example, the student annotation “Wild Type is seen as the natural, normal genes of an organism, whereas a mutation is a

Figure 2. The interactive decision tree.

type of change to these normal genes that can either provide a good or bad trait...” by itself appears to be highly engaged (Constructive). However, a course instructor, using this novel set of criteria, classified this annotation at a lower engagement level (Active) because the student paraphrased the textbook without adding significant insight.

Labeling cognitive engagement

Manually labeling students’ annotations according to their cognitive engagement is effort-intensive. The coding manual that was designed by Wang et al. [17] for this purpose includes 23 pages and includes definitions of each category and several example comments from a MOOC psychology course. The coding manual includes a decision tree in which nodes correspond to yes/no questions about annotations and leaves correspond to CE classifications.

Building on Wang et al.’s use of yes/no questions to distinguish between CE categories, we designed an “interactive decision-tree” application for classifying the CE of comments according to the CE taxonomy for annotation based forums. Figure 2 shows the adapted decision-tree for classifying CE. The application presents yes/no questions to the labeler, leading the labeler through the tree to the relevant CE label for a given comment.

In the process of manual CE scoring, we found that our readers often reflexively invoked ideas from the Bloom’s Taxonomy paradigm [1], to the assignment of CE. While Bloom’s and CE are complementary frameworks, they are functionally dif-

Course	# students	#annotations
AP50A	74	2,644
AP50B	72	2,786
BIS2A	716	2,212

Table 2. number of instances

ferent concepts in important ways. CE focuses on the degree to which students are involved with the material while Bloom presents a classification of levels of learning (Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating). An annotation that elaborates a student’s reasoning can exhibit engagement of type Constructive Reasoning even if the reasoning itself is incorrect and does not achieve the learning goals as framed by Bloom’s levels. In order to avoid this type of confusion, we made small modifications to the language used by Wang et al. in the text of the interactive decision tree.

We hypothesized that using the interactive application in lieu of an instruction manual would reduce the effort required from CE labelers without reducing their performance. Specifically, we expected that the interrater agreement between labelers would be similar to the one described by Wang et al. in their work. For both courses, we employed two labelers to label the same set of annotations using the interactive application. For the biology course, one labeler was a teaching assistant in the course, and the other was a 3rd year undergraduate (with 3.0 or above GPA) majoring in Biology (at a different institution than the one teaching the course). For the physics courses, both labelers were undergraduates (with 3.0 or above GPA) in their last year of Electrical Engineering studies who had both taken several physics courses. For both courses, we chose a set of annotations sampled from 4-8 threads in the forum, totaling 46 annotations for biology and 47 for physics.

We measured the interrater agreement between the labelers using the Kappa and Kendall-Tau statistics which are used to compare ordinal scales. The Kappa (and Kendall-Tau) agreement between the labelers using the interactive decision-tree was 0.692 (0.72) for biology and 0.683 (0.711) for physics. The Kappa statistics was reasonably close to the 0.72 Kappa agreement reported by Wang et al. [16], using the manual. For both courses, labelers reached a classification speed of about 50 annotations per hour. In comparison, for the physics course, two other student labelers using the manual to label 46 comments were able to label between 10-15 student annotations per hour. Although we cannot directly compare between these different settings, it is clear that using the interactive application was able to save time and effort to the labelers.

Data Analysis

Table 2 shows the data we collected for both courses. Annotations for each of the two physics courses were sampled from 17 chapters of reading while those for the biology course were sampled from a single chapter. Figure 3 shows the distribution of CE across the three courses. There are virtually no annotations labeled as O (Off Topic) and only a few labeled as A2 (Active General), the two lowest leveled categories of CE. By contrast, those studying engagement in non-annotated forums reports significantly higher Off Topic comments (See Discussion Section).

Figure 3. Distribution of CE for Physics and Biology courses

When comparing the CE distribution between courses, we can see that the ratio of annotations labeled as I (Interactive) in the Biology course (11.8%) is significantly lower ($\chi^2 = 220, p < 0.001$) than in the Physics courses (29.35% in AP50A and 24.01% in AP50B). Also, the ratio of annotations reflecting moderate CE (Active-targeted) in the Biology course (37.5%) was significantly higher ($\chi^2 = 197, p < 0.001$) than in the Physics courses (15.8% and 19.6%). We suggest several explanations for this difference.

One possible reason for this difference is the flipped classroom approach used in the Physics course. The forum was a main tool for communication between students while reading at home, which may have positively contributed to students generating more meaningful annotations. Second, the two courses included students from different populations. In Biology, the students represented 20 different majors and over 40 percent first generation college students, whereas in Physics, the students were engineering or premed students and about 15 percent first generation college students. A final possibility is simply cultural differences, where the labelers for Biology may have been consistently “harsher” in their labeling than the labelers for Physics, or possibly a real differences in thinking styles between the two domains.

Figure 4 shows the distribution of average CE per student in the AP50A Physics course (top) and Biology course (bottom). The distribution for the AP50B Physics course was similar to that of AP50A. We can see that in both courses, students’ exhibit a variety of CE behavior. Specifically, the average CE for most students in the two courses is Constructive-extending (C2) or above. Student CE exhibits a unimodal distribution with an average CE of 2.7 (where (-1) value is Off-topic and 4 value is Interactive) in AP50A, 2.56 in AP50B, and 2.145 in BIS2A. All courses exhibited the same standard deviation (1.08). For both courses, we found that the average CE of first posts (the comment that commences a thread) was lower than responses in all courses: 1.89 vs. 3.03 in Biology; 2.0 vs. 3.3 in Physics. This finding is also consistent with that of prior work [16] and can be attributed to the way CE is defined. First posts are commonly questions about the material, while responses engage with others by answering the questions or providing follow up questions.

Interestingly, while there is a degree of variance between the average CE across students, individual students mostly display

Figure 4. Distribution of CE per student. Biology (left), Physics (right).

a consistent CE level that does not change significantly over time. In the Physics courses, where the labeled annotations spanned the entire length of the course, the CE change for students in the AP50A and AP50B courses is -0.002 and 0.0005 respectively with a standard deviation of 0.188 and 0.1326 . This seems disappointing, as we might hope to see CE rise as students learn. But as we have discussed above, CE is distinct from understanding, and may not inherently rise as students learn. On reflection, if we believe that higher CE produces learning gains, it may be worth considering how to explicitly teach and encourage students to raise their level of CE.

We expected that students that are more active in NB (i.e., write more annotations) would also exhibit a higher average CE. Surprisingly, we found that students that write more annotations exhibit lower CE (Pearson correlation -0.33 , $p < 0.0079$ in both Physics courses), with similar results in Biology. The negative correlation is explained by the observation that students who write a lot, mostly generate questions, which is associated with lower CE. This explanation is verified in that the Pearson correlation between annotation count and question count were above 0.74 ($p < 0.0001$) for both physics courses.

Based on prior work we expected to see a relationship between CE behavior and course performance. In the biology course numerous assessments of learning were available that could be associated with CE and included total course grade, two independent mid-term exam scores, one final exam score, an average score computed from 9 questions that were directly related to the reading that were drawn from one midterm and the final exam. We calculated a Kendall rank correlation score between CE and each of these assessments for students having completed all of these assignments and found significant positive correlations for all assessments with CE: midterm 1 ($\tau=0.19$, $p=4.173e-10$), midterm 2 ($\tau=0.16$, $p=1.151e-07$), final exam ($\tau=0.17$, $p=1.077e-08$), total course grade ($\tau=0.21$, $p=5.932e-12$), and nine selected exam questions ($\tau=0.17$, $p=3.865e-07$). While these correlations do not implicate a causal association between CE and course performance they do seem to indicate that higher CE in the context of the forum is a behavior associated with greater success in the course. This may motivate us to explicitly *teach* students about higher-level CE in future courses. In the Physics course, exams included working in groups, so we cannot analyze the performance of individual students.

AUTOMATICALLY CLASSIFYING CE

In this section we describe the design of an automatic classifier for labeling CE using supervised learning. We used the data

Course	F1	Accuracy	Kappa	OCI
Physics (All)	0.75	0.774	0.626	0.31
Baseline+C	0.745	0.771	0.622	0.305
Baseline	0.69	0.698	0.526	0.411
Biology (All)	0.75	0.75	0.608	0.360
Baseline+C	0.73	0.72	0.55	0.39
Baseline	0.59	0.59	0.343	0.534

Table 3. Performance of Classification Algorithms

obtained for both the Biology BIS2A course, and another for the Physics AP50 courses (A and B). All the data reported in Table 2 was used in the classification. Our goal was to design a classifier that considers, in addition to features based on students' annotation, also the contextual features that exist in the reading and marked by the annotation. The classification labels were the same ones used by Wang et al. [17] which included labels for the three main on-task CE categories: Active, Constructive and Interactive.

Previous classifiers based on CE used bag-of-words models that did not consider the readings [17, 18]. We began with this model. We then augmented the model with *context* features relating the comment text to the text being annotated in the reading. Specifically, we compared the average cosine word similarity between the annotation body and its marked text, and the paragraph that surrounds the marked text. We also considered meta-data about the type of marked text, whether media, title, exercise, problem, solution, note, definition, list or other.

We considered another set of features, motivated by our analysis which suggested that an individual student's CE tends not to vary and that CE tends to be "sticky" by text region. We therefore considered student-specific features such as the number of annotations for the student, and the average CE for the student and the standard deviation of the CE for the annotations of this student. These CE-based features may not be available without significant labeling labor by a teacher; thus, we considered classification both with and without them.

We tested several classification models, including SVM, Decision Trees Random Forests and logistic regression. The classification model that obtained the best results on a validation set was a Random Forest classifier with a decision tree as the base classifier; we used the Gini impurity index as the criterion for deciding on which features to split. This is in agreement with the notion that the decision tree algorithms mirror human decision making [11, p. 315]. The best hyper-parameters for the classifier (tuned on the validation set) included 200 decision trees for training and a maximum of 16 features in each tree.

The results for classifying CE in biology and physics are shown in Table 3. For each class we compare three approaches. The *Baseline* approach uses a bag-of-words classifier that considers only information about the comment itself. *Baseline+C* incorporates the contextual features of the text being annotated and its metadata. The *All* results include features based on the student making the comment, such as their average CE on previous comments.

In addition to standard classification measures (F1, accuracy), we report two additional measures which are suited to predicting values on an ordinal scale: weighted Cohen's Kappa (higher is better), which considers the distance between categories when computing error functions, and the Ordinal Classification Index (lower is better), which adapts mean and absolute square error measures for ordinal dependent variables [2, 6]. All results were obtained using ten-fold cross-validation.

As shown by the table, for both courses, incorporating contextual features (Baseline+C) significantly improved the classification over the baseline. Adding the CE-based features increased performance slightly more. The improvement obtained from using the CE information was considerably more pronounced for the Biology course than the Physics course. This can be explained in that CE features were more informative for Biology than for Physics. Also, the Kappa measure achieved by the learning algorithm is similar to the one obtained from comparing between human experts. The main source of error for the learning algorithm in both courses was misclassification of the Active CE types (lower category) as Constructive CE types (higher category).

The most informative features for the biology course (based on information theoretic measures) were: whether the current annotation is a first post, the CE of the parent post (when the current annotation is not a first post), the average CE of previous annotations that overlap the marked area of the current annotation, the type of the marked text, the maximal and average cosine similarity between comments and the marked text, the average CE for this student, and the length (in words) of the annotation. The most important features for the physics course were: whether the current annotation is a first post, the number of question marks in the annotation, and the length of the annotation.

VISUALIZING CE

To date, CE has been used primarily by researchers studying learning [16, 18, 7]. By contrast, in this section, we show that CE labeling can help instructors visualize students' behavior. The design of the visualization tool was informed by a discussion with two of the teaching instructors in the course who were familiar with NB.

The instructors identified two key ways in which they would utilize CE to help understand their students' interaction with the text: (1) to better understand topics in the material that prompted high and/or low CE - it was initially thought that this might help to reveal where students were finding the material challenging in the text and that this might help to guide instructor interventions during the reading period itself; and (2) to use patterns of CE revealed prior to class as the basis for experimenting with lecture activities (no hypothesis was given about how high or low CE topics in the reading were to be treated in class, only that it would be interesting to try different class activities based on the prior knowledge of students' CE behavior).

Instructors indicated that because of the high comment volume per assignment (classes of 1,000 students generating 3,000 comments per reading) any CE reporting tool would need to:

Figure 5. CE heat map for NB.

(1) aggregate and display CE scores in ways that would not overwhelm the instructor; (2) for easy inspection, intuitively associate CE scores to the text from which the score was generated; (3) be visually intuitive to interpret.

The instructors also expressed a familiarity with and an appreciation for the intuitiveness of the existing heat-map tool in NB. This tool enables instructors to see document and page-level overviews of the commenting frequency on the text via variable-opacity color shading (high opacity indicates high comment density while low opacity indicates low comment density). Based on the familiarity of instructors with the heat-map, we hypothesized that an analogous heat-map tool could meet instructor needs by providing them with aggregate information regarding the location and quality of CE values throughout the text in a non-overwhelming, location-based, easy to use, and familiar format.

An example of the heat-map we developed is presented in Figure 5. The opacity of the color is defined by the number of annotations marking that area (like the classic NB heat map familiar to NB users), and the color of a text area is defined by the average CE of comments annotating that part of text. Darker shades (in red) are used to indicate annotation clusters with low value CEs, while lighter shades (in green and then yellow) are used to indicate annotation clusters with higher value. Hovering over the marked text displays to the user an average CE value for annotations that include the marked text. For example, in Figure 5, the area enumerated by "3" has a mid-range CE color (yellow-orange) and hovering over it shows that the CE level is 2.52, meaning that the average CE of the annotations there is between Active-targeted and Constructive-extending. Lastly, when selecting an individual annotation, the CE category of the annotation appears next to the student name. We believed that this visualization tool would meet the usability and information requirements described by the instructors.

User Study

In this section we report the results of a user study of the usefulness of the heat-map to instructors in the biology course. The user study was conducted with three instructors (two of whom participated in the initial discussion) and two teaching assistants. In this case all of the instructors had taught the course at least six times, with one also serving as the course coordinator. One of the teaching assistants served in that role for the course six times, and the other did so twice. All survey participants had actively use NB in the course before. Each of the instructors typically oversees 200-500 student classes. Since students' annotations were due by 11 pm the day before class, instructors usually had little time to read many individual student posts before class. By contrast, the teaching assistants, oversee much smaller groups of students (approx. 70 in groups of approx. 24) so keeping up with their section's comments - even with an 11 pm deadline - is less challenging.

The recruitment to the interview was voluntary. Staff motivation to participate was to use the findings of this research in their course and to help advance scientific research. One of the instructor subjects contributed to data collection for this study, however none of them were involved in the design of the CE heat map tool. All participants were given a tutorial explaining CE and its usage, as well as the use of the CE heat map. Before the interview, participants were asked to read a short explanation of CE definition. Then, they were presented with the new features the CE heat map adds to the NB interface and performed a simple training task.

The user study followed a semi-structured protocol [3] by which a pre-determined set of questions were used as an opportunity for the interviewer to explore further responses. All the interviews were conducted by the same interviewer.

As an important simplification, we evaluated the visualization using only the gold-standard comments that had been labeled by humans. This permitted us to avoid confounding the question of whether visualizing CE can help teachers with the errors in our classifier. In the future we aim to assess the effectiveness of our visualization when using automatically classified data.

The first part of the study included general questions about the use of NB in the class, how course staff select and read through annotations, and how the annotations affect their interaction with students in the class. The second part of the interview ask instructors about whether CE can be used to make inferences about students' engagement in the course, and in particular, whether they thought that using CE helps to locate areas in the textbook that students found interesting or difficult. The third part of the study focused on using the CE heat map find meaningful areas in the textbook. Participants were also asked to elaborate how they found these areas and why they thought they were meaningful. They were asked to compare the CE heat map to the default support tool in NB which highlights (undifferentiated) annotation density in the text (see Figure 1. Participants were also asked the speculate about their future use of the CE heat map and completed an official USE (Usefulness, Satisfaction, and Ease) questionnaire [12].

User Study Results

Assessing NB instructor behavior prior to CE heat map: In the first part of the study (questions about existing use of NB in the course), participants stated that their main activities are to search for unanswered student questions using a filter button (supplied in the NB GUI), as well as long threads (containing many comments) which they stated reflect students' misconceptions of the material. Some of the course topics that are relevant to these annotations are chosen to be further elaborated upon in class. Teaching assistants stated they review all annotations in their section in a sequential fashion. Both instructors and TA reported that they actively search for and commend (either in the forum or directly in class), students who generate "high quality" annotations which reflect understanding and insight of the material.

Assessing NB instructor behavior and affect when including the CE heat map: We present the user study results relating to the use of CE as an informative measure in the course. Here the participants needed to agree (on a scale of 1 to 5) with predefined sentences relating to the use of CE in class. All participants agreed that annotations exhibiting high CE reflect student interest and that that they indicate that students are "trying to figure something out". The average score for the statement "the CE heat map helps me find areas students will find interesting" was 4.0, and the average score for the statement "the CE heat map helps me find areas students will find interesting that I did not previously recognize." was 4.2. Participants associated student interest to topics in the course that they can relate with real life (e.g., Hamburger picture in the text book), visual information (seeing how the protein structure folds) and topics that are familiar to them from other classes. This agrees with Wang et al.'s finding in the Psychology course that course topics that relate to students' daily lives generate higher CE in their annotations [16]. One instructor added that while all annotations exhibiting high CE reflect student interest, other paragraphs in the textbook that reflect student interest may not exhibit high CE.

All of the participants connected the marked text in the reading material with the CE label of the relevant annotation. Some stated that questions to the reader that appear in the readings were more likely to exhibit high CE. All participants stated that definitions that appear in the readings were likely to exhibit low CE because the comments were paraphrasing the definition.

Interestingly, there was a consensus among the study participants that CE is not a good indicator for difficulty or confusion. According to all interviewees, the topic that students found most difficult is the subject of integrating pH and pKa. However, the instructors found that the comments within this region of the textbook exhibited both high and low CE. The statement "The heat map helps me find areas that students find difficult that I did not recognize" received an average score of 3.4. This finding reflects that the purpose and outcome of highlighting the CE taxonomy was not to categorize students' understanding of the material (like other cognitive theories, such as the Bloom's taxonomy) but rather, that it served correctly, to highlight varying levels of engagement.

Participants generally agreed that the heat-map helps instructors to choose where to provide feedback to students. The average score for the statement "The CE heat map helps me find threads in a specific area in the text where I would like to join the conversation or intervene." received an average value of 4.8 (out of 5). All of the participants stated that using the heat map will save them time when preparing for class.

They also consistently stated they would use the CE heat map to provide positive reinforcement (whether in the forum or in class) to students whose comments exhibit high CE. The statement "I will act when I see high CE" received an average value of 4.8. Instructors were less likely to intervene in a thread with low CE. The statement "I will act when I see low CE." received a mean of 3.4. Some instructors stated they would like to talk to students in class and find out the reason for the lack of engagement. Others stated that most annotations with low CEs paraphrase the text, and they would ignore such annotations.

There was an interesting difference between instructors and TAs in their use of the heat map. While instructors stated they would use the heat map to look for and respond to annotations in areas in the textbook exhibiting high CE, TAs stated they would use the heat map to prioritize their forum participation, first addressing annotations with low and high CE. One of the TAs stated that he would intervene when annotations exhibiting low CE refer to important definitions.

Visualizing CE also helped instructors to identify paragraphs in the textbook that may benefit from modification. Three of the participants explicitly said they would like to change parts of the textbook where low CE appears. All the participants mentioned that the paraphrasing phenomenon is a problem. All of these statements occurred after rating the statement "Low CE leads me to think there's a problem in that area." (average 4.2). One of the instructors suggested adding a few questions when he saw in the CE heat map that questions generate a higher level of CE.

The CE heat map was compared to the existing heat map that is supplied by NB, shown in Figure 1. This heat-map uses colors only to distinguish between the amount of annotation in different areas in the textbook, with no CE information. The users were asked to state "which heat map is more informative?" and "what are the extra benefits of using each heat map that the other one does not?". All of the interviews unanimously preferred the CE heat map to the traditional NB one.

Lastly, we describe in brief participants' response to the USE Questionnaire [12]. The questionnaire is constructed as a five-point Likert rating scale, where users are asked to rate agreement with the statements, ranging from strongly disagree to strongly agree. The CE heat map scored 5 out of 5 in all questions relating to "usefulness", "ease of learning", and "satisfaction"; it scored 4.5 out of 5 for questions relating to "ease of use". Participants said that they quickly learned how to use the CE heat map, that it is easy to remember how they use it, they intend to use it in their course, that they would recommend it to colleagues, and that they felt it would help them be more productive. In the topic of "ease of use", they

rated most questions with an average of 4.5. The statement "I can use the CE heat map without written instructions." received a lower score of 3.6.

DISCUSSION AND CONCLUSIONS

We have shown that measuring how deeply a student is thinking about and relating to course material, which has been shown to correlate with learning gains, is also a valuable pedagogical tool for teachers. Combining automatic classification and visualization of CE can give teachers insights about how students are thinking, and the ability to use these insights to improve class design, and to support beneficial discussions about the material. These insights augment other cognitive models of student learning that emphasize understanding (like Bloom's taxonomy). By anchoring comments into the text, annotation based forums, specifically the Nota Bene system, provide a useful framework with which to study and visualize CE to teachers. Models that reason about CE in the surrounding comments can improve the automatic classification of new comments, as compared to models that reason solely about contextual features.

Our work can be seen from two complementary perspectives. From the perspective of improving course readings, our approach of tying CE to specific text areas helps teachers understand where in the reading engagement is low, and focus their attention on bringing it up in that place, either by intervening in the discussion or by improving that specific reading material. From the perspective of helping teachers improve online discussion, anchoring that discussion in course readings through an annotation-based forum and heat map helps to coherently organize vast amounts of student discussion so that faculty can see a big picture of which topics require the most support and intervention. We also saw significantly less Off Topic labels in NB, which we attribute to the in-place structure of the forum. On the one hand, this is a positive result, since students' learning gains are positively correlated with their CE. On the other hand, we do not attribute a causal relationship between the two, therefore we cannot say that using NB in the course directly leads to better learning.

Our approach relies on the availability of several thousand labeled CE annotations. We used humans to obtain these labels, as our goal was to study the primal question of whether the CE measure can facilitate teachers' understanding of how students learn. In future work we will visualize CE labels to teachers obtained by a classifier, and verify that the error introduced by automatic classification does not destroy the utility of the CE visualization. Another topic for future work is to study the interplay between confusion and CE. For example, by using Emoji-generated hashtags which students use to tag affects like curiosity, confusion, and frustration [20].

A specific request from teachers is to improve the performance of the NB classifier by focusing specifically on identifying comments that paraphrase the readings. It would then be possible to filter out paraphrases if instructors want to ignore them anyway, or to give each student a "paraphrase score" so faculty could warn students doing it too much. Also, we would like to drop the assumption that comments are independent and using a classifier that biases towards comments on the same

text having similar CE. Another way to improve the classifier is to include dependencies between CE and location in text, following the observations that certain parts of the text exhibit high CE and certain parts exhibit low CE. Lastly, we note that the feature set that determine the CE varies across courses, hence it is necessary to train a different model for each course. In the NB domain, we were able to learn a single model for both courses without significantly reducing the results.

ACKNOWLEDGEMENTS

Thanks to Amy Zhang for helpful advice on the data analysis and the user study. This work was funded in part thanks to the Kamin fund by the Israeli Ministry of Trade and Industry and a grant from the National Science Foundation.

REFERENCES

1. Lorin W Anderson, David R Krathwohl, P Airasian, K Cruikshank, R Mayer, P Pintrich, J Rath, and M Wittrock. *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy*. New York. Longman Publishing.
2. Roger Bakeman and John M Gottman. 1997. *Observing interaction: An introduction to sequential analysis*. Cambridge university press.
3. AE Blandford. 2013. Semi-structured qualitative studies. Interaction Design Foundation.
4. 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).
5. Christopher Brooks, Omar Chavez, Jared Tritz, and Stephanie Teasley. 2015. Reducing selection bias in quasi-experimental educational studies. In *Proceedings of the fifth international conference on learning analytics and knowledge*. ACM, 295–299.
6. Jaime S Cardoso and Ricardo Sousa. 2011. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence* 25, 08 (2011), 1173–1195.
7. Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist* 49, 4 (2014), 219–243.
8. D Cormier and G Siemens. 2010. Through the open door: Open courses as research, learning, and engagement. *Educause* 45, 4 (2010), 30–39.
9. Gerardine DeSanctis, Anne-Laure Fayard, Michael Roach, and Lu Jiang. 2003. Learning in online forums. *European Management Journal* 21, 5 (2003), 565–577.
10. Ying-Hua Guan, Chin-Chung Tsai, and Fu-Kwun Hwang. 2006. Content analysis of online discussion on a senior-high-school discussion forum of a virtual physics laboratory. *Instructional Science* 34, 4 (01 Jul 2006), 279–311. DOI: <http://dx.doi.org/10.1007/s11251-005-3345-1>
11. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer. 315 pages.
12. Arnold M Lund. 2001. Measuring usability with the use questionnaire. *Usability interface* 8, 2 (2001), 3–6.
13. 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. *Physical Review Physics Education Research* 12, 2 (Dec. 2016), 19–12.
14. Dawn M. Poole. 2000. Student Participation in a Discussion-Oriented Online Course. *Journal of Research on Computing in Education* 33, 2 (2000), 162–177. DOI: <http://dx.doi.org/10.1080/08886504.2000.10782307>
15. Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 680–690.
16. 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*. ACM, 398–407.
17. 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.. In *The 8th International Conference on Educational Data Mining*.
18. Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2014. Linguistic Reflections of Student Engagement in Massive Open Online Courses.. In *Eighth International AAAI Conference on Weblogs and Social Media*.
19. Jian-Syuan Wong, Bart Pursel, Anna Divinsky, and Bernard J Jansen. 2015. An analysis of mooc discussion forum interactions from the most active users. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 452–457.
20. Amy X. Zhang, Michele Igo, Marc Facciotti, and David Karger. 2017. Using Student Annotated Hashtags and Emojis to Collect Nuanced Affective States. In *Proceedings of the Fourth Annual ACM Conference on Learning at Scale: Poster Publication (L@S '17)*. ACM.
21. Erping Zhu. 2006. Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science* 34, 6 (01 Nov 2006), 451. DOI: <http://dx.doi.org/10.1007/s11251-006-0004-0>
22. Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1883–1892.