New Methods for Confusion Detection in Course Forums: Student, Teacher, and Machine

Shay A. Geller, Kobi Gal, Avi Segal, Kamali Sripathi, Hyunsoo G. Kim, Marc T. Facciotti, Michele Igo, Nicholas Hoernle, and David Karger

***Abstract*—This article provides computational and rule-based approaches for detecting confusion that is expressed in students' comments in couse forums. To obtain reliable, ground truth data about which posts exhibit student confusion, we designed a decision tree that facilitates the manual labeling of forum posts by experts. However, manual labeling is costly in time and resources, which limits the amount of data that can be generated using this process. Our strategy for overcoming these limitations was to generate rules for detecting confusion based on student input via hashtags, which reflect the student's affective states. We show that the resulting rules closely align with the ground truth judgement of experts. We next applied these rules to datasets of students' forum posts in a large-scale biology course, thereby automatically generating thousands of labeled instances of "confused posts." Finally, the resulting dataset was used to train a machine learning model for detecting whether students' posts exhibit confusion in the absence of hashtags. In this task, the pretrained language model based on bidirectional encoder representation from transformers (BERT) was able to outperform traditional machine learning models for classifying confusion in posts. This model was also able to generalize and detect student confusion across different offerings of the same course. Ultimately, the use of pretrained language models of this type will provide teachers with better technologies for detecting and alleviating confusion in online discussion forums by leveraging the combined input of teachers and students.**

***Index Terms*—Educational technology, prediction methods, unsupervised learning.**

I. Introduction

BEHAVIORAL and emotional cues exhibited by students in the classroom communicate their level of interest, and whether they are confused, frustrated, or excited by the material the instructor is presenting [1]. Students' confusion when introduced to new material has been studied through lenses ranging from conceptual change to epistemic emotion (summarized in [2]). Although potentially viewed as a negative emotion by instructors and students alike [3], recent work suggests that confusion may either encourage or discourage student learning based on the other epistemic emotions that may accompany it [2], [4]. For instance, experiencing confusion, together with negative emotions (i.e., frustration and boredom), may impede student progress. When left unresolved, these emotions may lead to lower achievement, dropout, or academic dishonesty [5], [6]. However, when teachers recognize and respond to students who are experiencing negative emotions in time, students often return to an engaged state; thereby enhancing learning of the new material [2], [4], [5], [7], [8]. Thus, monitoring a student's emotional reaction to new material can enable faculty to provide interventions or other support to turn confusion into a positive learning experience.

As insight into students' epistemic emotions have evolved, so has the diversity of nontraditional learning environments and resources. Various technologies have increasingly moved traditional forms of course content delivery (e.g., readings, lectures, lecture notes, and research articles) online. Online content provides students with greater, more flexible, and sometimes more equitable access to course resources, particularly in cases where open educational resources substitute for fee-linked content. The move to online content delivery also provides the potential for creating more interactive content. Moreover, when appropriate technology is used, instructors have the opportunity to observe and understand how students are interacting with course content at scale in ways that were previously impossible with paper-based media. For example, an instructor who can get direct feedback about how their students are interacting with online course readings can identify topics within the readings, where students are experiencing confusion and other epistemic emotions and intervene productively on the students' behalf.

The opportunity for instructor intervention with student confusion during online learning is a powerful tool for individualizing learning within the context of large-scale classrooms.

Manuscript received June 8, 2020; revised July 7, 2021 and October 2, 2021; accepted October 16, 2021. Date of publication October 27, 2021; date of current version January 17, 2022. This work was supported in part by the National Science Foundation under Grant 1915724. *(Corresponding author: Kobi Gal.)*

Shay A. Geller and Avi Segal are with the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel (e-mail: gelleral@post.bgu.ac.il; avise@post.bgu.ac.il).

Kobi Gal is with the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel, and also with the School of Informatics, University of Edinburgh, EH8 9AB Edinburgh, U.K. (e-mail: kobig@bgu.ac.il).

Kamali Sripathi, Hyunsoo G. Kim, Marc T. Facciotti, and Michele Igo are with the Genome Center and the Department of Biomedical Engineering, University of California, Davis, Davis, CA 95616 USA (e-mail: ksripathi@ucdavis.edu; hyunsookim@ucdavis.edu; mtfacciotti@ucdavis.edu; mmigo@ucdavis.edu).

Nicholas Hoernle is with the School of Informatics, University of Edinburgh, EH8 9AB Edinburgh, U.K. (e-mail: n.s.hoernle@sms.ed.ac.uk).

David Karger is with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02142 USA (e-mail: karger@mit.edu).

Digital Object Identifier 10.1109/TLT.2021.3123266

However, key steps for effective instructor intervention are the accurate detection of student confusion and the ability to link this confusion to a particular location within the course material. Instructors can then use this information to guide their interactions with the students, develop online materials to help students with particularly problematic topics, and inform better course design. We propose that the students' self-reported affective states in the form of hashtags can be used to identify when, and on what content, the students are expressing confusion. Furthermore, when hashtags are not present, as observed in most online student forums, we still aim to provide useful insights about confusion to the course instructors by harnessing the power of machine learning (ML) to automatically classify the text in students' posts.

In the absence of hashtags, a natural way to identify confused posts in a forum is to rely on course staff or trained experts for the manual labeling of the posts. For example, the Stanford MOOCPosts dataset contains thousands of posts that were manually labeled on a Likert-type scale of confusion by nine colleagues [9]. However, as we worked with our course experts, we discovered that there was not a universally agreed upon definition of confusion (Section IV-A). Based on these discussions, we decided to introduce more rigour into the labeling process by designing a labeling tree that leads experts toward more consistent label assignments.

Such manually generated labels can be used to train ML classifiers, which can identify confused posts in a dataset based on linguistic cues and the vocabulary used in the students' posts [10], [11]. However, this approach imposes significant burden on the course staff, who in many cases are required to do the hand-labeling of the posts.

We propose an alternative approach that infers confusion with course material directly from the students' use of hashtags. In this article, we investigate the following research questions.

1) Research Question 1 (RQ1): Does the introduction of a structured labeling tree for guiding experts result in higher agreement, compared to using a Likert-type scale measure?
2) Research Question 2 (RQ2): Do students' hashtags agree with experts' judgments about what constitutes confused posts?
3) Research Question 3 (RQ3): Can students' hashtags be used to inform automatic classification of confused posts for situations in which hashtags do not exist? Can these models generalize to other course terms?

To answer these research questions, we suggest three new approaches for identifying confusion in this setting [3].

The first approach facilitates manual labeling using experts (course teaching staff). We designed a labeling tree that helps the course teaching staff distinguish between the different types of confusion found in student posts. The labels in the tree reflect a taxonomy of different types of confusion, such as seeking information when students are aware they lack it, and making incorrect statements about course material without being aware they are incorrect. We found that using the labeling tree significantly improved interreliability agreement between experts when compared to traditional methods of collecting opinions, such as the Likert-type scale [5], [10]. Although this tree-based method is optimal with respect to accuracy in detecting "confused posts," it is resource intensive in terms of both time and money. This led us to explore other approaches that might generate a good proxy for teachers' judgments about the type of confusion exhibited by students in posts. We use the decision-tree approach to generate hundreds of ground truth labels about whether students' posts express confusion. We then use these labels to evaluate the reliability and reproducibility of the second and third less resource intensive approaches.

The second approach classifies posts exhibiting confusion based on students' use of hashtags in the posts. In many forums, student are given the option to select from a set of predefined hashtags in their posts to convey opinions, ideas, and emotions in a similar way to many social media platforms [12]. By comparing student labeling using hashtags to labeling by experts using the decision tree approach, we show that a naive labeling rule that checks for the presence of the predefined #confused hashtag is a sufficient, but not necessary condition for inferring confusion in a student's post as defined by the course teaching staff. This discovery motivated a new labeling rule, which considers additional hashtags that convey questioning and help-seeking behaviors, which may also be indicative of confusion. We show that this new labeling rule is more aligned with teachers' ground truth judgement of confusion in the students' posts. We use this rule to generate thousands of proxy labels of confusion in students' posts in a fast and efficient way. The third approach uses ML classifiers for detecting confusion in students' posts that can be applied in the absence of student self-reported hashtags. We leveraged the proxy labels from the vast number of posts with students' self-reported hashtags to create a model trained on a dataset of posts labeled using the second approach. Experts then evaluated the validity of the ground truth labels generated by this model on posts without hashtags using the labeling tree we described in the first approach.

We showed that a deep learning-based model, which has been pretrained on general text corpora and used only on the raw text of the post as input, was able to outperform traditional ML models that rely on feature engineering. Moreover, this model was able to generalize between different offerings of the course, making it a good candidate model to be used in practice.

This article shows that by making use of the students' self-annotated posts, we can augment existing models for confusion detection and can inform the development of automatic confusion detection systems to support teachers' understanding of how students comprehend course readings. It also demonstrates the value of including the "human-in-the-loop" in the design of analytical tools for supporting online learning. Providing students with a natural and familiar way to convey affective states improved the performance of ML models for detecting confusion, even for cases where hashtags were not available. Moreover, by directly involving teachers in the process of validating our rules, we were able to arrive at nontrivial conclusions for what constitutes confusion in online forums.

This article significantly extends a prior conference publication [13] in several ways. First, we provide a richer definition of confusion in students' posts that is informed by the educational literature. We provide a thorough discussion of this literature in the introduction and related work. Second, we design a structured way to collect labels from experts using the labeling tree. Third, we provide new ML models for detecting confusion that outperform those in the earlier conference article, and are able to generalize well across different instances of the same course.

II. Related Work

This article relates to educational literature that aims to define and understand students' confusion when engaging with course material, and to literature in learning technologies that addresses the problem of automatically identifying students' confusion in online discussion forums.

A. Confusion in Students' Learning

Confusion has been defined in a variety of fields [2]. In the realm of education, Perkins [14] has defined students' confusion as the opposite of "understanding," which is a student's ability to grasp content or to retain and actively apply their knowledge. In cognitive psychology, Piaget [15] historically framed confusion through a psychological lens as a disequilibrium: Students may experience confusion when they have preconceived conceptual models but are unable to assimilate new information that is coherent for their models. Finally, in the field of conceptual change, confusion has been identified as a potential side effect when students are confronted with the ways new knowledge differs from their existing conceptions [16].

In this article, we selected the definition of confusion used in the field of epistemic emotion [3]. Our rationale is that in contrast to some of the lenses mentioned above, work in epistemic emotion will allow us to take into account both instructor and student perspectives of confusion. These dual perspectives are essential to our work in two ways: From a practical standpoint, the instructor perspective is needed as the reference to train our predictive model. From an education standpoint, providing students with efficient instructor feedback in online learning environments is crucial to students' success in these novel environments. We, therefore, adopt Plaut's framework of confusion [3], due to the rich detail it provides and its ability to include both the student and instructor perspectives. Plaut proposes the presence of multiple facets of confusion. Out of the four facets that his work presents, two are relevant to this work: *Type* and *cause*. *Type* relates to the aspects of learning that students find confusing, such as the learning material itself, the teacher's presentation and explanation of the material, or the students' understanding of the teacher's expectations regarding the course material. *Cause* relates to the underlying reasons for the confusion. Examples of cause can be unclear directions from the teacher, students' misconceptions about the course content/goals, insufficient preparation by the student (i.e., did not do homework), being "tuned out" (daydreaming) in class, or reading material that is insufficient or too complicated for their level of understanding. One way to distinguish between the different *types* of confusion is to answer the question, "What is the student confused by?" The way to distinguish between the different *causes* of confusion to answer the question, "What is the reason for the student's confusion?" These definitions relate to our work in Section VI-C3, where we show that certain types and causes of confusion are more difficult than others to identify in forum posts.

Importantly, Plaut [3] also shows that teachers and students may have different perspectives of what constitutes confusion. In practice, we experienced the implications of teachers' varied perspectives when we asked a number of teachers and course material experts to label forum posts as displaying confusion. This led to the work in Section IV-A in which we describe how we, in conjunction with the teachers, designed a labeling tree to increase the consistency of assigning the confusion label to individual students' posts.

Lodge *et al.* [17] and Arguel *et al.* [18] presented novel and nonintuitive results, which indicate that confusion may, in fact, be beneficial for learning. Some works even embrace confusion and study the effect of how difficulties might contribute to learning in online environments. Research areas, like desirable difficulties [19], productive failure [20], impasse-driven learning [21], cognitive disequilibrium [22], and discovery-based environments [23], purposely inject confusion into the learning process in different ways and test its effect on student learning.

However, when confusion is not resolved in a timely manner, it can have negative effects on a student's learning. When students are unable to resolve their confusion, it triggers frustration that will eventually lead to boredom and disengagement [8]. Beck *et al.* [24] also proposed the concept of "wheel-spinning," where confused students fail in mastering a skill in a timely manner. This may lead to them giving up and never mastering the skill due to associated negative emotions which may lead to further disengagement. Shute *et al.* [25] studied the effect of persistent confusion in educational games and they suggest that targeted interventions can help students overcome this confusion. Indeed, they stress the importance of fast detection and early identification of confusion.

B. Manual Detection of Confusion in Forum Posts

The most comprehensive dataset containing posts that are manually labeled for confusion is the Stanford MOOCPost dataset [10]. The dataset contains 29 604 annotated forum posts from 11 Stanford University public online courses. Nine expert colleagues labeled six emotions conveyed in posts, including confusion. They used a seven-point Likert-type scale to measure the level of confusion [10], [11], [26], [27]. A different approach was used by Geller *et al.* [13], which relied on a single expert from the course's teaching staff, who identified the presence of confusion in forum posts. The course staff member assigned confusion labels to posts by answering the question, "To what extent does the following post exhibit the student's confusion on the relevant reading material?" and

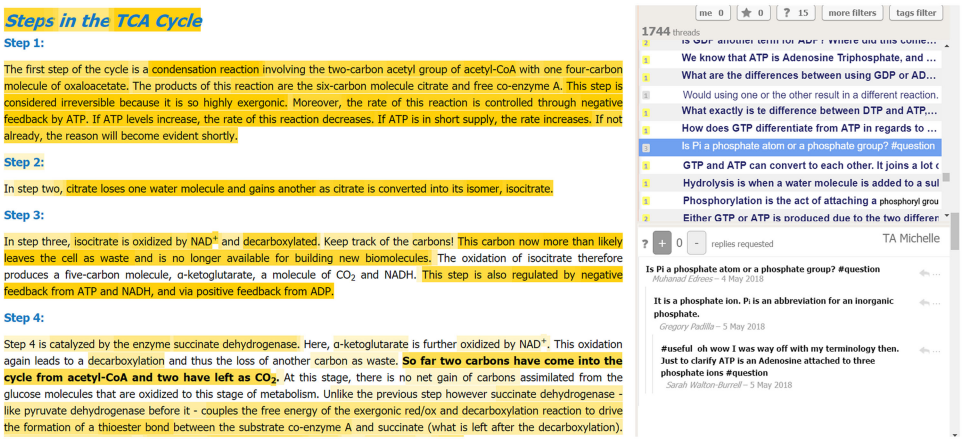


Fig. 1. Nota Bene UI. Left: Course reading material. Highlighted text represents portions of text with an annotation. Right: A list of annotations.

used a four-point Likert-type scale to measure the level of confusion with 1 being "no confusion" and 4 being "very confused."

Yang *et al.* [5] used human labelers from MTurk to label thousands of students' posts in Algebra and Microeconomics courses. The crowd-sourced labelers were asked to judge the level of confusion exhibited in the posts on a four-point Likert scale. We extend this work by providing a more nuanced definition of confusion, one informed by pedagogical research. While the MTurk approach is highly scalable, it has the limitation that it does not involve human experts.

Zhang *et al.* [28] used student-provided hashtags in Nota Bene (NB) to label curiosity and confusion. They assumed a one-to-one mapping from hashtags to affective states. However, they did not evaluate their approach with human experts.

C. Automated Labeling of Confusion in Forum Posts

Most work on ML models for confusion detection has been based on the Stanford MOOCPost dataset [5], [10], [11], [26], [27]. We list some of the main approaches below. Agrawal *et al.* [10] used a bag-of-words approach to represent posts, and included metadata information about the post and the sentiment of the post. Their approach used a logistic regression model to predict confusion on several of the Stanford MOOC forums, achieving best performance on forums that include technical discussions on topics such as statistics and economics. Zeng *et al.* [11] also considered community-related features of the post, such as the number of reads and the number of upvotes of each post. They showed that the accuracy of the classifier monotonically increases with confusion level of the post (measured on a Likert-type scale of 1–7). Wei *et al.* [26] used convolution and recurrent neural networks for the confusion detection task. Their pretrained, deep learning-based language models achieved state-of-the-art results on several common prediction tasks [29]. Models of this type are pretrained on vast amounts of text data to create rich universal language representations that can be used instead of handcrafted features in classification tasks [30], [31]. Clavié and Gal [27] used bidirectional encoder representations from transformers (BERT) to achieve state-of-the-art results on confusion detection in the MOOCPost dataset.

Here, we examine the efficacy of using BERT for confusion detection in the Nota Bene forum and show that it can outperform models relying on hand-defined features used by Geller *et al.* [13].

III. Learning Environment

Our empirical methodology uses data from the NB platform, an open-source collaborative online annotation tool [32]. Course content (PDF, HTML, video files) is uploaded to the NB website by instructors. Students can annotate the content by highlighting a passage in the document (called "the marked text") and then add a post by typing into a text field that appears in the margin. These annotations may be used to make a comment or to ask questions about the content. Classmates are encouraged to reply to other student's comments and to answer any posted questions. NB annotations are organized into threads, which consist of a starting comment or question followed by all the replies made by other students to the initial annotation or to the subsequent replies. NB is used in hundreds of university courses and there are more than 40 000 registered student users.

Fig. 1 shows the NB interface for a biology course that is part of our empirical methodology. On the left is a section of the course textbook, which describes the "TCA Cycle," a key metabolic pathway discussed extensively in introductory biology courses. The reading material is augmented by annotations from students and faculty, which appear as expandable discussions on the right-hand-side panel. Annotations are anchored to particular locations in the document based on the content highlighted by the student, who made the starting comment of the thread. One can explore the annotations by looking through the text and selecting a highlighted section of interest, which also brings up the corresponding annotation(s) on the right panel and/or by scrolling through the list of annotations on the right panel, where selecting one will bring the user to the corresponding highlighted text. The reading on the left embeds a heat map that highlights areas in the document associated with many (dark yellow) and few (light yellow) annotations.

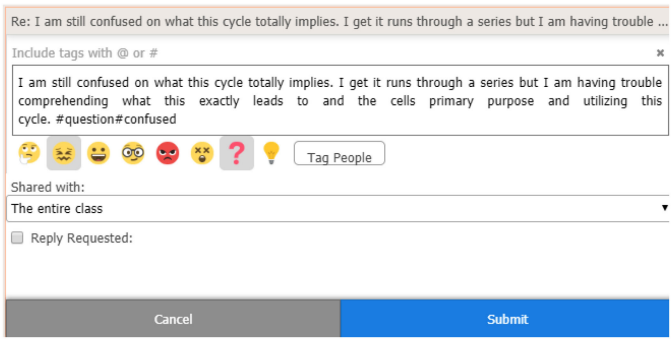


Fig. 2. Nota Bene hashtags GUI including a post with the hashtags, and relevant emojis clicked (gray background surrounds them).

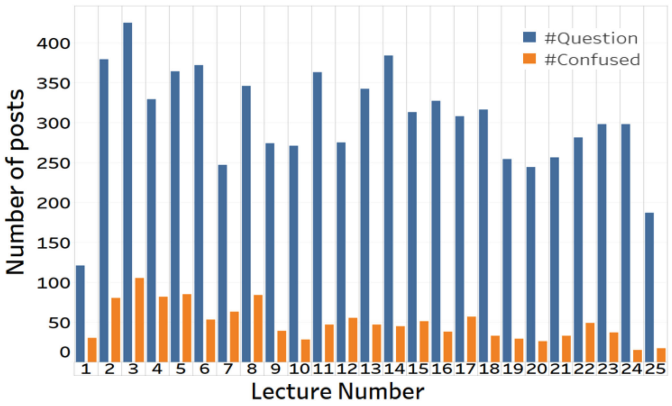


Fig. 3. Distribution of #confused and #question hashtags over lectures in the summer term.

TABLE I
HASHTAGS, THEIR ASSOCIATED EMOJIS, AND THE COUNTS OF THEIR USE IN THE SUMMER TERM

Hashtag	Count	Emoji	Hashtag	Count	Emoji
#confused	1228		#help	420	
#curious	6595		#question	7574	
#interested	9237		#useful	10 866	
#idea	9673		#frustrated	65	

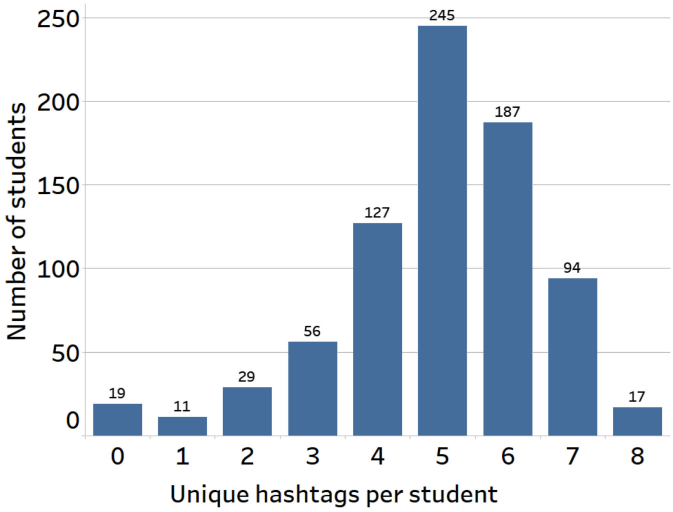


Fig. 4. Histogram of unique hashtag use by students in the summer term.

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. The application of NB in educational settings has shown to be beneficial; one study showed that there was a positive correlation between students who engaged in high-level discussions and learning gains in exams that tested their conceptual understanding [33].

Bis2A is a general biology course required for all life sciences majors, many social sciences majors, and bioengineering students at the University of California, Davis.

Our dataset consists of two terms of Bis2A, in summer and winter of 2018. Each term consisted of 25 lectures. Reading assignments for each lecture were uploaded to NB and students were required to provide "three meaningful posts" for each reading assignment in NB before the lecture. The purpose was to encourage active participation in forum discussions and to provide a way for students to obtain answers to their questions before lecture.

The NB interface also allows students to generate pictorial representations (emojis) of their affective state using hashtags and text, (e.g., #confused, #useful, #frustrated) as shown in Fig. 2. Table I shows the eight possible hashtags, including the frequency of their usage by the students in the summer term. (The distribution of hashtags in the winter term was similar, hence for brevity we display results for summer term only in this section.) Students received additional credit for including at least one hashtag in at least one of their posts per lecture. Another feature of NB is that posts can be filtered by hashtags, allowing instructors to navigate to posts that display specific emotions.

In total, out of 58 811 posts in the summer term, 40 842 posts (70%) contained at least one hashtag. Similarly, out of 70 360 posts in the winter term, 35 016 (50%) contained at least one hashtag. These proportions are well above the minimal requirement required from students. This suggests that students may perceive intrinsic value from associating hashtags with their comments, supporting calls for providing students with opportunities for self-assessment [34].

Fig. 3 is a histogram showing the number of posts displaying #confused and #question hashtags across the different lectures in the summer term. As shown by the figure, students' use of hashtags varied across the different lectures. The information derived from their use of hashtags can contain useful insights for the course staff. For example, lecture three, which *a priori* was considered to be easy by the course staff, had the highest number of posts containing the hashtag #confused (105 posts) and #question (425 posts), suggesting that students found this material difficult. In contrast, lecture ten, which *a priori* was considered to be difficult by the teaching staff, had relatively few number of posts containing the hashtag #confused (28 posts) and #question (271 posts).

Students' use of hashtags varied widely, reflecting prior results showing the varied use of hashtags in popular social

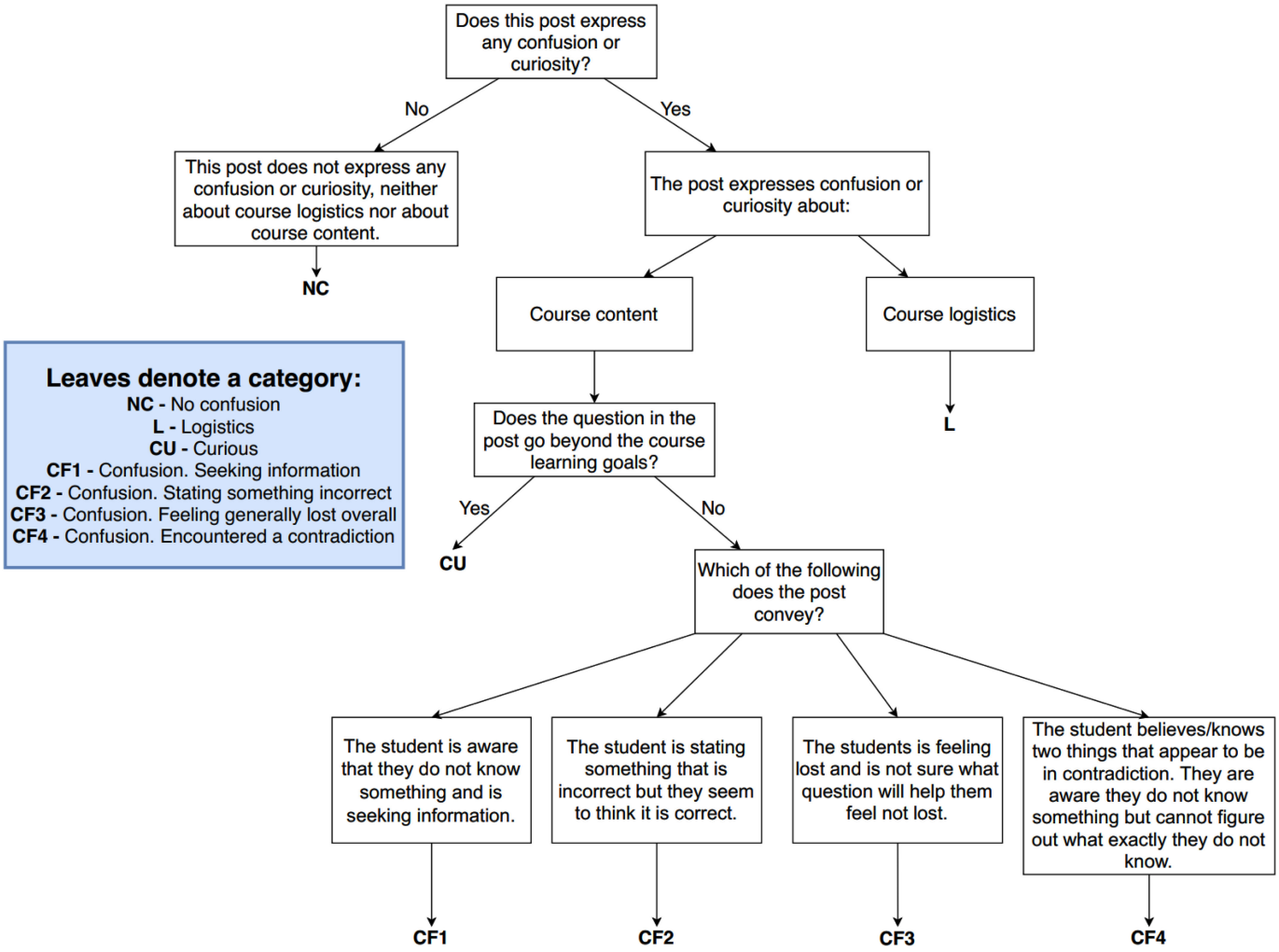


Fig. 5. Labeling tree for confusion.

media platforms [35]. Fig. 4 shows a histogram of the number of hashtags used at least once by students across all of the lectures in the summer term. As shown by the figure, the majority of the students (more than 85%) made use of four or more of the eight available hashtags with a small minority (2.2%) using all of the hashtags. In general, students contributed well above the minimal number of required hashtags for each lecture. Only a small minority of the students did not include any hashtag in their posts (2.5%).

IV. RQ1. Interlabeler Agreement When Using Labeling Tree

A. Methodology for RQ1

In this section, we describe the methodology for addressing research question 1. We design a labeling tree for collecting expert labels for confusion in NB.

Previous works [10], [13] used a Likert-type scale to obtain confusion labels from experts. We asked two experts from the Bis2A course staff to label over 200 posts using a Likert-type scale. Comparison of the label assignments revealed low agreement ratios between the two experts (Kappa measure of. 39 over 200 posts). This identified the need for a more structured approach to labeling confusion that reflects the taxonomy of different types of confusion from the pedagogical literature [3]. To this end, we developed a method that takes into account the different aspects of confusion by using a labeling tree, presented in Fig. 5. Other works have demonstrated the benefits of using labeling trees to facilitate analysis of students' forum behavior [36], [37].

The labeling tree is composed of a series of questions to be answered by the experts (course staff members) and it directs them to choosing a label that reflects the type of confusion exhibited by the post. Based on discussions with the course staff and manual analysis of students' posts, we identified several types of confusion that are commonly expressed by students: Feeling lost about course material, drawing an incorrect conclusion, asserting there are contradictions in the text, and seeking information about course material or logistics. We also distinguished questions about material the students are expected to learn, from questions that go beyond the course learning goals. Table II presents a complete list of label categories and an example of a post that exhibits each category.

TABLE II
DEFINITION AND EXAMPLE FOR EACH LEAF LABEL OF THE LABELING TREE

Tree Label	Definition	Example
NC	No confusion	"This can be remembered easily because the prefix 'mei' means to reduce!"
CU—Curiosity	Posts contain statements or questions that go beyond the course learning goals. The same question can be labeled as confusion or curious, depending on whether the student was expected to know the concept at that particular stage in the course or not.	"I wonder what other types of exposure can affect formation of different mistakes in DNA. Is it just radiation?"
L—Course Logistics	Posts relate to course technicalities like assignments, exams, or course requirements. For example, questions such as: "Do we need to know _____ for the exam?" Our experts seldom label these types of questions as confusion. Though this indeed is a type of confusion related to expectations for the students, we would like to separate them from cases where confusion is more directly related to misunderstanding the course content.	"If the main focus is to give a general view on how energy and matter are relevant in biological systems, will specific cases of energy such as in mitochondrial interactions and glycolysis be less important to know? Or will those still be required in future lessons?"
CF1—Confusion—Seeking Info	The student is aware that they do not know something and is seeking information. This usually involves asking questions or wondering about specific portions of the text, or soliciting other students' opinions. This label captures most of the confused posts.	"Is this saying that there is no specific start and end of the ETC?"
CF2—Confusion—Wrong	The student is stating something that is incorrect, but they seem to think it is correct.	"Yes, unlike the bonds in ATP, hydrogen bonds require energy to be broken."
CF3—Confusion—Lost	The student is feeling lost and is not sure what question will help them feel not lost.	"I searched some info but I can't be sure that is correct. It says that in higher plants, the cyclic photophosphorylation helps with the condition that ATP is not sufficient."
CF4—Confusion—Contradiction	The student encounters two things that appear to be in contradiction. They are aware they do not know something, but cannot figure out what exactly they do not know.	"How would you change the nucleotide sequence? I thought the strength of the promoter does not get altered? I thought activators and repressors in some sense adjust to 'regulate' the strong/weak promoters."

B. Results for RQ1

We begin by describing interlabeler agreement, when experts used the labeling tree from Section IV-A. Then we describe the agreement between the two labeling rules $R1$ and $R2$ and the experts.

To evaluate the labeling tree, we randomly picked 300 posts with no hashtags from each term of the Bis2A course, making 600 posts in total. We denote posts without hashtags as *raw posts*. Three experts from the course staff labeled the 300 posts from the summer course term independently using the labeling tree of Fig. 5. Experts labeled all their posts before comparing their work. They achieved an intraclass correlation coefficient of 0.86, indicating high agreement among the experts. We further presented the 300 posts from the winter course to two of the three experts. They independently used the labeling tree to classify the confusion in the posts and achieved a Kappa score of 0.90.

We can, thus, answer RQ1 affirmatively, in that the Kappa score obtained with the labeling tree (0.90) was significantly higher than that obtained with the Likert-type scale approach (0.39). All of the experts stated that using the labeling tree significantly facilitated the labeling process in terms of coherence.

V. RQ2. AGREEMENT BETWEEN STUDENTS' SELF-REPORTED HASHTAGS AND EXPERTS

A. Methodology for RQ2

In this section, we describe the methodology for addressing research questions 2. Using students' self-reported hashtags, we suggest labeling rules for selecting confused posts.

To annotate posts for confusion using students' self-reported hashtags, a natural labeling rule to consider is a one-to-one mapping from #confused hashtags to confused posts. An example of such a post can be seen below.

> *During lecture, the differences [*sic*] between hydrogen bonds and ionic bonds were discusssed [*sic*]. I'm still having difficulty understanding the differences. #confused.*

However, some posts that show confusion do not include #confused hashtags. For example, consider the following post.

> *I'm a little bit confused about the difference of [*sic*] atom and elements. In chemistry, I call the elements "atom," but I don't really understand the difference and assumed that they are the same. How do we distinguish atoms and elements? #question*

This student is expressing confusion about the reading, yet is not using the #confused hashtag.

TABLE III
COSINE SIMILARITY BETWEEN THE EMBEDDING FOR THE #CONFUSED HASHTAG AND THE OTHER HASHTAG EMBEDDINGS

Hashtag	Cosine Similarity
#help	**0.987**
#question	**0.982**
#curious	0.973
#frustrated	0.907
#interested	0.89
#idea	0.885
#useful	0.881

The closer the value is to 1, the more similar it is to #confused.

Interestingly, the use of the #confused hashtag makes up only 3% (1228) and 11% (4178) of the total number of the 40 842 and 35 016 posts with hashtags in the summer and winter terms, respectively. Thus, it is either the case that 1) the set of all confused posts is a superset that includes posts labeled with a #confused hashtag or that 2) the amount of confusion displayed in this dataset is significantly lower than the amount noted by previous work [11], [28]. The above examples clearly show that the naive labeling rule, which only uses the #confused hashtag, is inadequate for detecting all of the confused posts, leading us to conclude that the first explanation is the correct one. We, therefore, need a new rule that goes beyond the #confused hashtag to identify all the confused posts in student forums.

To find such a labeling rule for confusion in this dataset, we use a computational method to investigate whether the hashtags are used in similar contexts (with similar vocabulary in their posts). We follow a method similar to that introduced by Eisner *et al.* [38], which uses word embeddings to create vector representations for Twitter emojis. Similar to Wei *et al.* [26], who trained their word embedding model on the posts from the Stanford MOOCPost dataset, we trained word embeddings using the NB posts. We then calculated the cosine similarity between the embedding vector of the #confused and the other hashtags to define a new labeling rule that takes into consideration other hashtags rather than just #confused.

To compute the embeddings, we trained a Skip-gram version of the Word2Vec model [39], with 50 hidden neurons in the embedding layer, on the text from all of the posts on the system. We used the model to compute an average vector to represent each post by averaging the word level embeddings of the text in the post. We further averaged all post level embeddings for each hashtag. If hashtag a is cosine similar to hashtag b, this means that posts labeled with both a and b contain semantically related words. Table III shows the cosine similarity between the vector embedding of each hashtag to the vector embedding of the #confused hashtag . As shown by the table, both the #help and #question hashtags show the highest values for cosine similarity to the #confused hashtag (in bold script).

This suggests that our new labeling rule should consider #help and #question hashtags in addition to the #confused hashtag.

Based on the cosine similarity scores, we grouped students' hashtags into two groups. A group containing hashtags that reflect a need for assistance (#question, #help, #confused) and a group containing hashtags that reflect interest in the material (#idea, #useful, #interested, #curious). We ignored the appearance of #frustrated due to its low occurrence in the data (it appears in less than 0.02% of all posts). This analysis led us to consider and compare the validity of the following two rules for detecting confused posts, the naive labeling rule (R1) and our data-informed rule (R2).

R1: This rule declares a post to be confused if and only if it contains a #confused hashtag.

R2: This rule declares a post to be confused if and only if it contains *at least* one hashtag from the assistance hashtag group, and *none* of the hashtags from the interest group. Thus, R2 is a superset of R1.

We hypothesize that rule $R1$ is a sufficient condition for defining confused posts, but not a necessary condition. That is, a human expert could identify confused posts that satisfy rule $R2$, but not rule $R1$.

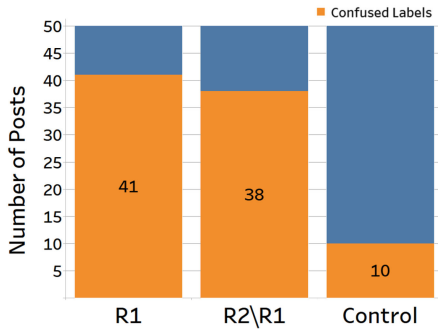


Fig. 6. Agreement between expert and posts in groups $R1$, $R2 \setminus R1$, and control group.

B. Results for RQ2

To test the agreement between expert labels and our two labeling rules that are based on students' self-reported hashtags, we formed three groups of posts from the set of posts in the summer term of Bis2A. Each group contained 150 posts with every post containing at least one hashtag. The first group included posts satisfying rule $R1$. The second group included posts satisfying rule $R2$, but not $R1$. The third group was a control that included posts that satisfy neither $R1$ nor $R2$. Due to the high interreliability agreement between experts on the 300 posts described in Section IV-B, the labels for these 150 posts were generated by a single expert.

We say that the expert agrees with a rule for a given post if the expert labels that post as exhibiting a degree of confusion (CF1, CF2, CF3, or CF4 labels of the tree). Fig. 6 shows the number of agreed instances between the expert and the posts in each group (orange bars). We refer to the posts in the groups above using the rule that generated them: 1) Group $R1$; 2) group $R2 \setminus R1$; and 3) control group.

Statistical significance was obtained using Z tests with Bonferroni adjusted α level of 0.05. We used the normal approximation to the Binomial distribution as the sample

size was large in all three cases. From the figure, we can conclude that 1) there was significantly more agreement between the expert and the posts in both groups $R1$ and $R2\backslash R1$ than the posts in the control group; and 2) there was more agreement between the expert and the posts in group $R1$ than in the group $R2\backslash R1$, but this difference was not significant. The Kappa agreement between the expert and the posts in the $R1$ group was 0.28. However, the Kappa agreement between the expert and posts in the $R2$ group (which includes both the $R1$ group and $R2\backslash R1$ group) was 0.55. This supports our premise that the labeling by experts is more likely to agree when the posts are labeled using the $R2$ rule. Moreover, it is important to note that the 38 posts labeled correctly as confused under the rule $R2$, would have been missed by the rule $R1$.

These results confirm our hypothesis that while $R1$ may be the most important indication for confused posts, it is incomplete, and adding posts that satisfy rule $R2$ will improve the detection of confusion in this dataset. Thus, we have answered research question 2 in the affirmative.

VI. RQ3. Automatic Classification of Confused Posts

A. Methodology for RQ3

In this section, we address RQ3, which asks if can we predict students' confusion in raw posts (that do not contain hashtags) using ML. We first discuss in detail the datasets that were used for training and testing the different ML classifiers. We then introduce the feature engineering for preprocessing the data before classification. Finally, we present the ML models that were tested.

1) Datasets: The training set for the models consisted of posts from summer and winter terms of Bis2A that were classified as confused using the $R2$ rule (hence referred to as $R2$-labeled datasets). The test set for the models consisted of raw posts from summer and winter terms of Bis2A that had been labeled by the experts (hence referred to as expert-labeled datasets). Any of the labels CF1, CF2, CF3, or CF4 in the tree were considered to be "confused," and the other labels (NC, CU, L) were considered to be "nonconfused."

For the summer term, where we used three experts, we discovered that the majority of the posts were not labeled as "confused" by the experts. Although 64 were labeled as "confused," 236 posts were labeled as "nonconfused." For the winter term, where we used two experts, we kept only those instances, where the two experts agreed on the post label. This resulted with 288 posts, with 52 labeled "confused," and 236 labeled "nonconfused."

2) Feature Design: We applied standard preprocessing to the raw text. The preprocessing steps included 1) the removal of words that appeared less than five times (to reduce vocabulary size and avoid spelling mistakes), and 2) the retaining of stop-words and punctuation (which can provide information about confusion).

A number of natural language techniques [40] were used to extract features from the raw text and the context of a post. The context of a post from NB consists of three parts: The "post-context" (the content in the post), the "highlighted-context" (the selected text in the reading which the post relates to), and the "paragraph-context" (the entire paragraph in the reading material that contains the highlighted context). We used the following feature families.

3) General Purpose Textual Features—Bigrams: We extracted word-level bigrams, which are general-purpose textual features [41], [42], from the post-context. The extracted word-level bigrams' tokens counts were normalized using TF–IDF scores for each token [43].

4) Linguistic Features: Linguistic inquiry and word count (LIWC) [44] is a software package that analyzes text and extracts multiple categories of features that represent emotions, attention, sentiment, etc. This tool has been widely used in similar studies of affective states [5], [45], [46].

For all of the contexts, we used the following features [47]: "we," "i," "you," "shehe," "ipron," "affect," "posemo," "negemo," "negate," "nonflu," "insight," "assent," "adverb," "certain," "dicrep," "certain," "compare," "quant," and "differ."

Certain other LIWC features were only relevant to the paragraph-context. Therefore, for only the paragraph-context, we extracted the "see," "hear," and "feel" LIWC features.

The highlighted area of the text associated with most confused posts is located in paragraphs close to a figure/plot or to some example or explanation. We, therefore, further extract features from the paragraph-context that can help us identify which of these paragraphs contained the words "figure," "example," and "consider." We also counted the number of words that only contain numbers (i.e., 1920) and words that contain both characters and numbers (i.e., H2O, 100 μm).

Similar to [5], we also tried to identify sentences that explicitly express some confusion. To accomplish this, we also searched for the presence of the following statements in the post-context: "I am confused," "I was stuck," and "I am struggling with."

5) Sentence Complexity: Prior work on text readability has used features that capture the complexity of sentences [48]. Complex sentences within the text could indicate difficult paragraphs, which may lead to more posts displaying student confusion. We considered a sentence to be a sequence of words separated by a period or a question mark. The sentence complexity features were: Average number of words, characters, nouns, and adjectives per sentence. Moreover, we included the number of exclamation points in these features.

6) Direction and Action Items: The readings contain descriptors that guide understanding toward concepts in the course. Usually, such sentences start with a verb like "Consider" or "Find." They also may contain other listing connectives words like "above," "then," "first," "second," "finally," etc. These features are Boolean variables that indicate whether or not the paragraph-context contains one of these words.

7) Question Features: Questions are commonly used to express confusion [5]. To identify the questions, we extracted the following features over the post-context: The number of

TABLE IV
WITHIN-TERM RESULTS

Summer

Model	Precision	Recall	F1	Accuracy	PR AUC
LR+B	0.83	0.53	0.65	0.87	0.74
LR+B+CW	0.79	0.72	0.75	**0.9**	0.74
LR+B+F	0.83	0.53	0.65	0.88	0.7
LR+B+F+CW	0.78	0.66	0.71	0.88	0.7
BERT	**0.9**	0.56	0.69	0.89	**0.85**
BERT+D	0.74	**0.8**	**0.77**	0.89	0.83

Winter

Model	Precision	Recall	F1	Accuracy	PR AUC
LR+B	0.74	0.54	0.62	0.88	0.71
LR+B+CW	0.65	0.83	0.73	0.88	0.72
LR+B+F	0.77	0.58	0.66	0.89	0.73
LR+B+F+CW	0.67	0.83	0.74	0.9	0.74
BERT	**0.81**	0.75	**0.78**	**0.92**	**0.78**
BERT+D	0.69	**0.87**	0.77	0.91	0.74

Models trained and tested on summer term (top) and winter term (bottom).

question marks, whether a post contains a sentence that starts with a modal verb (i.e., can, could, be) or a question word (i.e., what, how, who).

B. Models

We experimented with several common ML classifiers. The training set for each model consisted of all posts with hashtags in the Bis2A course (summer and winter) and were labeled as expressing confusion by the $R2$ rule. We compared between logistic regression, k-nearest neighbors, support vector machine, decision trees, random forests, and XGBOOST. Logistic regression with L1 regularization consistently outperformed the other models using tenfold cross validation.

We compared the performance of the logistic regression model, with the custom features defined in Section VI-A2, with a general purpose deep learning model commonly used for NLP tasks. To this end, we used the BERT model, a pretrained, transformer-based language model that is commonly used in state-of-the-art natural language tasks [29]. This model, which is pretrained on huge general text corpus from books and Wikipedia articles, was fine-tuned on the Bis2A posts. The fine-tuning of the BERT model involved two additional learning epochs, with a maximal sentence length of 512 words, using the "bert-base-uncased" configuration using the HuggingFace open-source library [49].

We also compared performance of the different models with and without using inverse class-weight balancing and downsampling to correct for the class imbalance.

C. Results for RQ3

We test the performance of the classifiers on a number of configurations of training and testing.

1) *Within-Term:* We train and test the classifiers on $R2$-labeled posts from one of the two terms, and test the classifiers on the expert-labeled posts from the same term.
2) *Between-Terms:* We test the cross-term generalization of the classifiers by training on the $R2$-labeled posts from one of the two terms, but testing on expert-labeled posts from the other term.

All classifier parameters were hypertuned using tenfold cross-validation on only the $R2$-labeled posts. The metrics used for comparing the classifiers' performances include precision, recall, F1, predictive accuracy, and precision-recall (PR) area under the curve (AUC). Davis and Goadrich [50] show that the receiver operating characteristic curves can present overoptimistic results with skewed data. Thus, it is preferable to use PR AUC as a reliable alternative [51]–[53]. As the model becomes both more correct and more confident about its predictions, the higher the PR AUC number becomes, trending toward 1 for a perfect classifier.

In the following sections, we present the performance results of six classifiers.

1) LR+B: A logistic regression model that is only trained on the word level bigrams features.
2) LR+B+CW: The LR+B classifier with inverse class-weight rebalancing applied.
3) LR+B+F: A logistic regression model that is trained on the word-level bigrams and the domain specific features.
4) LR+B+F+CW: The LR+B+F classifier with inverse class-weight rebalancing applied.
5) BERT: A BERT model that is fined-tuned directly on the posts' raw text.
6) BERT+D: The BERT model with downsampling rebalancing applied.

1) Classification of Confusion Within Terms: Table IV compares the classifier performance on the summer (top) and winter (bottom) expert-labeled datasets, respectively.

For the summer term datasets, the BERT models (shown in bold) outperform the other models in all metrics except accuracy. This is also true for the winter term dataset, where the BERT models (shown in bold) even outperforms the others on accuracy. While the accuracy metrics across all models are high (≥ 0.87), this is partly due to the imbalanced test data where "not-confused" is the majority class. The PR AUC metric shows that the BERT models are much more confident in their predictions than the feature engineering approaches. However, some of the logistic regression models, like LR+B +CW, achieve competitive results to the BERT models in the F1 metric, and there was no single BERT model that was best in all metrics.

The results also show that class rebalancing techniques on the training data result in an increase in recall and F1 across all models, but a decrease in precision.

Interestingly, adding the domain-specific features did not contribute to better model performance on the summer term. Notably, the LR+B+CW model outperforms LR+B+F+CW in every metric. This is not consistent with the results from the winter term, where the domain features did lead to an increase in model performance across all metrics. This result illustrates that a complex relationship can exist between the contributions of the domain specific features across different terms.

2) Classification of Confusion Between Terms: Table V compare the classifier performance when training on one term and testing on the other term. The table presents only the

TABLE V
BETWEEN-TERMS RESULTS

Summer

Model	Precision	Recall	F1	Accuracy	PR AUC
LR+B	↓ 0.05	↓ 0.09	↓ 0.09	↓ 0.03	↓ 0.05
LR+B+CW	↓ 0.06	↓ 0.02	↓ 0.04	↓ 0.02	↓ 0.04
LR+B+F	↓ 0.05	↓ 0.09	↓ 0.09	↓ 0.03	–
LR+B+F+CW	↓ 0.03	↑ 0.04	↑ 0.02	↑ 0.01	↓ 0.01
BERT	↓ 0.04	–	↓ 0.01	↓ 0.01	↓ 0.03
BERT+D	↑ 0.04	↑ 0.03	↑ 0.03	↑ 0.02	↑ 0.01

Winter

Model	Precision	Recall	F1	Accuracy	PR AUC
LR+B	↑ 0.05	↑ 0.19	↑ 0.14	↑ 0.03	↑ 0.05
LR+B+CW	↑ 0. 9	↓ 0.02	↑ 0.04	↑ 0.03	↑ 0.05
LR+B+F	↑ 0.02	↑ 0.05	↑ 0.04	↑ 0.01	↑ 0.06
LR+B+F+CW	↑ 0.08	↑ 0.02	↑ 0.05	↑ 0.02	↑ 0.03
BERT	↑ 0.05	↓ 0.04	–	–	↑ 0.04
BERT+D	↑ 0.02	–	↑ 0.01	–	↑ 0.05

Models trained and tested on different terms. results on summer term (top) and winter term (bottom). ↑, ↓ represent an increase or decrease over same term's results from Table IV.

difference in performance between the within-term and the between-term settings. Importantly, the results shows that performance did not decrease for all metrics and all models, when the models were trained on the summer dataset and tested on the winter dataset.

Interestingly, the same universal increase in performance is not observed when models trained on the winter dataset were tested on the summer dataset; in fact, most classifiers' performance decreased across all metrics. An exception to this rule is that the BERT+D does improve upon its winter predictions when trained on the summer dataset. A possible reason for this is that the summer training set was larger than the winter training set (40 000 posts versus 35 000 posts). Despite this fall in performance, the decreased performance observed with the models is not more than 9% on the F1 measure. The results show some evidence that the classifiers generalize across the contexts of terms.

3) Error Analysis: We study in detail the predictions of the BERT+D model, which was the best performing model over both terms. We compare the predictions made by this model to the expert labels collected in the within-term setting. Table VI presents results of the number of misclassified posts for summer and winter terms. The table shows that BERT-D performed best on the two classes for most labels, achieving an almost perfect prediction performance for the "No Confusion" (NC) and "Confusion—Seeking Info" ($CF1$)) labels.

However, the BERT-D model does not perform as well with the other labels, namely, "Confusion—Wrong" ($CF2$), "Confusion—Lost" ($CF3$), "Confusion—Contradiction" ($CF4$), and "Curious" (CU), which occur infrequently in the datasets. Although the sample size is low, we see a trend that the misclassification rate on these posts is high, with a vast majority of these posts getting misclassified in the summer and winter, respectively.

We highlight two categories that the model performed particularly poorly on, the $CF2$ and CU labels. The "Confusion—Wrong" ($CF2$) label denotes a post, where the student makes a factual error without realizing it. Since the

TABLE VI
ERROR ANALYSIS OF BERT+D MODEL OVER EXPERT-LABELED POSTS ON THE SUMMER AND WINTER TERMS

	Summer			Winter		
	Total	misclassified	%	Total	misclassified	%
CF1	51	3	6	45	0	0
CF2	7	7	1	7	7	1
CF3	3	3	1	0	0	–
CF4	3	0	0	0	0	–
NC	221	5	2	219	1	.4
L	2	2	1	3	3	1
CU	13	11	85	14	14	1

students are unaware of their misconceptions, they do not use the #confused, #question, or #help labels that would be used to train the classifier under $R2$. Moreover, any classifier hoping to correctly identify this form of confusion would need a conceptual understanding of the content. This understanding, which is embodied by the experts during their labeling, goes beyond the ability of the natural language techniques that we applied. It is also important to note that the only labeling system to correctly identify these posts is that which depends on the expert labelers. Even the $R2$ labeling rule, which depends on the students' self-reported confusion, cannot identify this form of confusion in the dataset.

Second, the "Curious" (CU) label was used by the experts to identify statements that go beyond the course learning goals.

As with the $CF2$ label, we can neither expect the students nor the classifier to identify this label, because it requires a thorough understanding of the biology course curriculum.

VII. DISCUSSION

The results of our three research questions highlight the difficulty that arises in labeling confusion in forum posts. First, one might assume that the students' own perspectives about confusion would align with that of experts. However, our error analysis (Section VI-C3) shows there exist examples where this is not the case. Second, teachers and domain experts also represent a source for labeling confusion, but due to the varied interpretations of confusion, it is difficult to reach a consensus on the labels (Section IV-A). We expand on both of these insights and discuss the potential implications therein.

A. Labeling Confusion in Forum Posts

The results of the three research questions presented in this work reflect the different perspectives about students' confusion [3] as it is exhibited in online forums. We found that using the labeling tree helped teachers align their understanding of the degree of confusion in students' posts. The adoption of the tree resulted in a better agreement among the teachers who completed the labeling. We consider the teachers to represent the ground truth for what posts are truly confused, as they are able to identify, in particular, the $CF2$ (Confusion—Wrong) and CU (Curiosity) labels. It is important to note that even the students cannot identify these two labels themselves as they 1) do not always realize their misconception, and 2) are not always confident about the bounds of the scope of the course.

While we motivate the practical use of the labeling tree, we emphasize that was not meant to provide a holistic definition for confusion. Rather, the tree was designed with the input of the Bis2A teaching staff, for the goal of reducing space for varied interpretations. For example, a student might be confused about course logistics. However, one teacher might not deem this as useful to the goal of understanding confusion in the reading material. This teacher might not label that post as confused, whereas a different teacher might rely more on the mere presence of confusion, and thus, label it differently. By separating these various labels out, we have reduced the possibility that these multiple interpretations might occur and have thereby increased the alignment among the teachers' labels.

Due to a lack of precise definitions for the labels, there exists varied use of self-reported annotation among students. In Section V-B, we show that the trivial rule for labeling confusion (one which merely considers the use of #confused) is insufficient for identifying a large number of posts that truly do reflect confusion. We, therefore, present a better rule, one which considers more than just the #confused label and we show that this new rule better captures the set of all confused posts. We argue that students' self-reported hashtags should be used to label confusion in the posts, as we cannot rely on the staff to hand-annotate the thousands of posts on a forum. While the students have difficulty identifying the $CF2$ (Confusion—Wrong) and CU (Curiosity) types of confusion, we show that the scalability of our approach outweighs this downfall.

B. Automatically Detecting Confusion in Forum Posts

We have discussed the value of using students' self-reported labels for identifying confusion in their posts. However, many posts do not contain such self-reported labels. For example, in our datasets, approximately $1/3$ of all posts did not contain a hashtag, and this was the case even though the students were explicitly asked to include them. It is clear that relying solely on students self-reported hashtags is insufficient for labeling all the posts in a dataset. However, we show that an automated classifier can be trained to predict confusion on posts that do not contain hashtags. This classifier was trained on labels derived from the students' self-reported hashtags, but it was evaluated on the labels that were provided by the teaching staff who used the labeling tree. This automated classification is extremely helpful, as it will identify the presence of confusion, even when students do not label their own posts with a hashtag.

The overall performance of the BERT-D model (F1 measure of 0.77) makes it a good candidate model to be used in practice. Despite this, we identified that certain confusion types could not be reliably detected. In particular, the $CF2$ and CU labels were missed by the classifiers. However, this was not unexpected. The two reasons for the poor performance on these labels are 1) the classifiers were trained on the data from students self-reporting, which would also be expected to miss these labels, and 2) the classifiers do not understand the semantics of the course material or the syllabus and, therefore, cannot possibly identify these two labels.

C. Pedagogical Implications

A few implications come from this work that address the way we should label students posts in online discussion for confusion and highlight the limitations of the current ML models for this task.

Our work highlights the importance of building ML models for detecting confusion that are informed by the pedagogical literature. The ability to separate the multiple types of confusion using the tree structure is an important expressive power, which is lacking in other labeling methods that only rely on Likert-type scale methods. Therefore, we conclude that given a student's post, an expert course staff member who uses the labeling tree is better qualified to label the post for confusion than the student. Even though we showed a good mapping from students' self-reported hashtags to experts' labels, some types of confusion, that may be rare but are expected, seem to be only identifiable by experts. Despite this, we show that the students' use of self-reported hashtags were very useful in training the ML models.

Finally, the course staff stated that labeling posts using the tree helped them to better understand how the students interact and understand the reading material. For example, the teachers were interested in understanding what in the reading material led students to write incorrect statements ($CF2$ labels). Because these comments are associated with highlighted regions in the text, it would be relatively easy for the teachers to add clarifications or examples to paragraphs containing the highlighted regions, which could help resolve these misconceptions. Thus, the labeling tree, combined with students' posts in the discussion forums, can be used as a pedagogical tool for teachers to obtain insights and resolve issues in the reading material.

D. Theoretical Implications

As summarized in [2], confusion has many definitions and implications in different schools of thought. Historical definitions characterize it as negative or undesirable [14], [15], whereas conceptual change literature views it as a necessary step in students' corrections of their misconceptions [16]. Recent work on confusion's epistemic properties highlight possibilities for both negative [5], [6] and positive [2], [4] consequences. We adopted this more nuanced view in the current work: The easier it is for instructors to detect student confusion, especially in novel online learning environments, the greater the chance for instructors to guide students toward positive confusion-related outcomes. To that end, we applied Plaut's [3] framework to identify common types of confusion in students' comments, allowing us to define the four types of confusion (CF1–CF4) in our decision tree.

Our results, specifically our between-terms predictions, also highlight the multifaceted nature of confusion before any academic consequences (good or bad) can be observed. The strongest classifier, the BERT model, performed best on "Confusion—Seeking Information" (CF1). This category is the most straightforward depiction of a student's awareness of their own cognitive disequilibrium, because students appear to be asking very straightforward questions. The second

straightforward example of disequilibrium is "Confusion—Contradiction," which occurs very infrequently. This infrequency may be due to the amount and sophistication of a student's knowledge. In order to fall in the CF4 category, students must both identify what they do and do not know, and compare these two domains to specify their questions. As previously stated, this category may be outside the scope of classifier prediction because of the in-depth content knowledge required.

Our remaining Confusion categories may provide an expansion of the epistemic definition of confusion. "Confusion—Wrong" (CF2) and "Confusion—Lost" (CF3) both do not contain explicit questions (hence, likely why students typically do not use hashtags for these comments). Both of these categories represent "timepoints" in student understanding before the students are aware of their own cognitive disequilibrium. Conversely, the instructors are aware of this mismatch in student thought, as evidenced by instructor creation of our CF2 and CF3 categories. Thus, although their identification is practically challenging, these two categories would likely be the most useful for instructors, because it would allow them to intervene before students ignore or forget these gaps in their own understanding. In addition, these two categories emphasize the role of the instructor as a resource in helping students identify, and eventually resolve, mistakes in new knowledge incorporation.

VIII. Conclusion

In this article, we study three research questions related to detecting confusion in student discussion forums.

The first research question discuss methods for labeling confusion by experts. We show that confusion is prone to multiple interpretations, even by course staff members, and simply using Likert-type methods for labeling confusion can lead to poor agreement between the experts. We propose a labeling tree that consists of a series of questions to be answered by experts (course staff members), and it directs them to choose a label that reflects the type of confusion exhibited by the post. This tree was shown to produce consistent labels between experts by facilitating the labeling process in terms of coherence.

The second research question asks whether students' self-reported hashtags agrees with experts' judgement about what constitutes a confused post. We show that the #confused hashtag is sufficient, but not a necessary condition for inferring confusion in a student's post. We learned an embedding vector for each hashtag, and show that the #help and #question are semantically similar to the #confused hashtag. This led to the design of two labeling rules, the naive one that uses only #confused to signify a confused post, and our data-informed rule that also incorporates the #help and #question. We show that the latter is more aligned with experts' judgement.

In the third research question, we ask whether we can harness students' self-reported hashtags to inform automatic classifier for detecting students' confusion in posts in which hashtags do not exist. We experiment with models that rely on hand-designed features, as well as with state-of-the-art pretrained language models (BERT) that accepts only the raw text of the post as an input. We show that classifiers trained using the hashtag-labeled posts perform well on expert-labeled posts and that the BERT model outperformed other models on the task. We also show that the models can generalize between terms, training on data from one term and predicting on another. Finally, we present an error analysis of the BERT model, which shows that some types of confusion are harder to detect than others. This illustrates the importance of the hashtags as valuable training data, and the fine-grained confusion labels obtained when experts used the labeling tree as a way to understand the models' true blind spots.

We also note the limitations of our approach in capturing only some aspects of students' confusions and not capturing other forms of confusion [16]. In future work, we wish to extend the ML models to improve prediction of the less obvious confusion types such as CF2 (wrong statement is not aware of). Also, we wish to design a confusion "heatmap" of students' posts with a color scheme of the degree of their expressed confusion. This will provide teachers with a high-level picture of the level of confusion throughout the reading material. Also, we wish to study how feedback from this heatmap can guide teachers in their design of the course.

Acknowledgment

The authors would like to thank A. Zhang for her contributions to the earlier version of this paper, which appeared in LAK-2020.

References

[1] R. T.Pekrun, W.Goetz Titz, and R. P.Perry, "Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research," *Educ. Psychol.*, vol. 37, no. 2, pp. 91–105, 2002.

[2] S. B. D'Mello, R. Lehman Pekrun, and A. Graesser, "Confusion can be beneficial for learning," *Learn. Instruct.*, vol. 29, pp. 153–170, 2014.

[3] S. Plaut, "'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*, vol. 36, no. 4, pp. 391–421, 2006.

[4] M. M. T. Rodrigo, R. S. Baker, and J. Q. Nabos, "The relationships between sequences of affective states and learner achievement," in *Proc. 18th Int. Conf. Comput. Educ.*, 2010, pp. 56–60.

[5] D. M. Yang, I. Wen, R. Howley Kraut, and C. Rose, "Exploring the effect of confusion in discussion forums of massive open online courses," in *Proc. 2nd ACM Conf. Learn. Scale*, 2015, pp. 121–130.

[6] G. J. A. Alexandron, Z. Ruipérez-Valiente, P. J. Chen Muñoz-Merino, and D. E. Pritchard, "Copying, scale: Using harvesting accounts for collecting correct answers in a MOOC," *Comput. Educ.*, vol. 108, pp. 96–114, 2017.

[7] B. M. Grawemeyer, W. Mavrikis Holmes, and S. Gutierrez-Santos, "Adapting feedback types according to students' affective states," in *Proc. 17th Int. Conf. Artif. Intell. Educ.*, 2015, pp. 586–590.

[8] S.DMello and A. Graesser, "Dynamics of affective states during complex learning," *Learn. Instruct.*, vol. 22, no. 2, pp. 145–157, 2012.

[9] A. Agrawal and A. Paepcke, "The stanford MOOCPosts dataset," Stanford Univ., Stanford, CA, USA, 2014.

[10] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke, "YouEDU: Addressing confusion in MOOC discussion forums by recommending instructional video clips," in *Proc. Int. Conf. Educ. Data Mining*, 2015.

[11] Z. Zeng, S. Chaturvedi, and S. Bhat, "Learner affect through the looking glass: Characterization and detection of confusion in online courses.," in *Proc. 10th Int. Conf. Educ. Data Mining*, 2017.

[12] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, 2015.

[13] S. A. *et al.*, "#confused and beyond: Detecting confusion in course forums using students' hashtags," in *Proc. 10th Int. Conf. Learn. Analytics Knowl.*, 2020, pp. 589–594.

[14] D. N. Perkins, *Smart Schools: Better Thinking and Learning for Every Child*. New York, NY, USA : Free Press, 1992.
[15] J. Piaget, *The Psychology of Intelligence*. Evanston, IL, USA: Routledge, 2005.
[16] M. T. Chi, "Three types of conceptual change: Belief revision, mental model transformation, and categorical shift," in *Proc. Int. Handbook Res. Conceptual Change*, 2008, pp. 61–82.
[17] J. M. G. Lodge, L. Kennedy, A. L. Arguel, and M. Pachman, "Understanding difficulties and resulting confusion in learning: An integrative review," in *Proc. Front. Educ.*, 2018, p. 49. [Online]. Available: https://www.frontiersin.org/article/10.3389/feduc.2018.00049
[18] A. L. Arguel, O. V. Lockyer, J. M. Lipp Lodge, and G. Kennedy, "Inside out: Detecting learners' confusion to improve interactive digital learning environments," *J. Educ. Comput. Res.*, vol. 55, no. 4, pp. 526–551, 2017.
[19] E. L. Bjork and R. A. Bjork, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," in *Psychology and the Real World: Essays Illustrating Fundamental, Contributions to Society.* vol. 2, New York, NY, USA: Worth, 2011, pp. 55–64.
[20] M. Kapur, "Productive failure," *Cogn. Instruct.*, vol. 26, no. 3, pp. 379–424, 2008.
[21] K. VanLehn, *Toward a Theory of Impasse-Driven Learning*. New York, NY, USA: Springer, 1988, pp. 19–41.
[22] A. C. S. Graesser, B. A. Lu, E. Olde Cooper-Pye, and S. Whitten, "Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down," *Memory Cogn.*, vol. 33, no. 7, pp. 1235–1247, 2005.
[23] L. P. J. Alfieri, N. J. Brooks Aldrich, and H. R. Tenenbaum, "Does discovery-based instruction enhance learning?," *J. Educ. Psychol.*, vol. 103, no. 1, p. 1–18, 2011.
[24] J. E. Beck and Y. Gong, "Wheel-spinning: Students who fail to master a skill," in *Proc. 16th Int. Conf. Artif. Intell. Educ.*, Memphis, TN, USA, Jul. 2013, pp. 431–440.
[25] V. J. Shute *et al.*, "Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game," *Comput. Educ.*, vol. 86, pp. 224–235, 2015.
[26] X. H. Wei, L. Lin Yang, and Y. Yu, "A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification," *Information*, vol. 8, no. 3, 2017, Art. no. 92.
[27] B. Clavié and K. Gal, "EduBERT: Pretrained deep language models for learning analytics," 2019, *arXiv:1912.00690*.
[28] A. X. M. Zhang, M. Igo Facciotti, and D. Karger, "Using student annotated hashtags and emojis to collect nuanced affective states," in *Proc. 4th ACM Conf. Learn. Scale*, Cambridge, MA, USA, Apr. 2017, pp. 319–322.
[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[30] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Proc. 29th Conf. Neural Inf. Process. Syst.*, 2015, pp. 3079–3087.
[31] M. E. Peters *et al.*, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.
[32] S. Zyto, D. Karger, M. Ackerman, and S. Mahajan, "Successful classroom deployment of a social document annotation system," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Austin, TX, USA, May 2012, pp. 1883–1892.
[33] K. S. Miller, D. Zyto, J. Karger Yoo, and E. Mazur, "Analysis of student engagement in an online annotation system in the context of a flipped introductory physics class," *Phys. Rev. Phys. Educ. Res.*, vol. 12, no. 2, 2016, Art. no. 0 20143.
[34] J. A. Ross, "The reliability, validity, and utility of self-assessment," *Practical Assessment, Res. Eval.*, vol. 11, 2006, Art. no. 10.
[35] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using Twitter hashtags and smileys," in *Proc. 23rd Int. Conf. Comput. Linguistics: Posters*, Beijing, China, Aug. 2010, pp. 241–249.
[36] E. K. Yogev, D. Gal, M. T. Karger Facciotti, and M. Igo, "Classifying and visualizing students' cognitive engagement in course readings," in *Proc. 5th ACM Conf. Learn. at Scale*, London, U.K., Jun. 2018, Art. no. 52.
[37] X. M. Wang Wen, and C. P. Rosé, "Towards triggering higher-order thinking behaviors in MOOCs," in *Proc. 6th Int. Conf. Learn. Analytics Knowl., Ser*, Apr. 2016, pp. 398–407.
[38] B. T. Eisner, I. Rocktäschel, M. Augenstein Bošnjak, and S. Riedel, "emoji2vec: Learning emoji representations from their description," 2016, *arXiv:1609.08359*.
[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[40] S. M. J. Vijayarani Ilamathi, and M. Nithya, "Preprocessing techniques for text mining–An overview," *Int. J. Comput. Sci. Commun. Netw.*, vol. 5, no. 1, pp. 7–16, 2015.
[41] J. Ramos *et al.*, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instructional Conf. Mach. Learn.*, vol. 242, Washington D.C, USA, Aug. 2003, pp. 133–142.
[42] L. S. Jensen and T. Martinez, "Improving text classification by using conceptual and contextual features," in *Proc. Workshop Text Mining 6th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2000, pp. 101–102.
[43] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.
[44] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count: LIWC*. vol. 71, Mahway, NJ, USA: Lawrence Erlbaum Assoc., 2001.
[45] O. A. Almatrafi Johri, and H. Rangwala, "Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums," *Comput. Educ.*, vol. 118, pp. 1–9, 2018.
[46] T. Atapattu, K. Falkner, M. Thilakaratne, L. Sivaneasharajah, and R. Jayashanka, "An identification of learners' confusion through language and discourse analysis," 2019, *arXiv:1903.03286*.
[47] J. Pennebaker, M. Francis, and R. Booth, *Linguistic Inquiry and Word Count: (LIWC).*Mahway, NJ, USA: Lawrence Erlbaum Assoc., Jan. 1999.
[48] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2008, pp. 186–195.
[49] T. Wolf *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.
[50] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 233–240.
[51] M. Craven and J. Bockhorst, "Markov networks for detecting overlapping elements in sequence data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, Dec. 2005, pp. 193–200.
[52] R. R. Bunescu Ge *et al.*, "Comparative experiments on learning information extractors for proteins and their interactions," *Artif. Intell. Med.*, vol. 33, no. 2, pp. 139–155, 2005.
[53] S. Kok and P. Domingos, "Learning the structure of Markov logic networks," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 441–448.

Shay A. Geller received the B.Sc. degree in computer science, and the M.Sc. degree in artificial intelligence and data science from the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 2018 and 2020, respectively.

He is currently a Data Scientist with HiredScore, Tel Aviv, Israel.

Kobi Gal received the Ph.D. degree in computer science from Harvard University, Cambridge, MA, USA, in 2006.

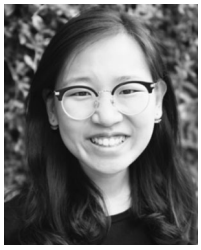
He is currently an Associate Professor with the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, and a Reader with the School of Informatics, University of Edinburgh, Edinburgh, U.K.

Avi Segal received the Ph.D. degree in artificial intelligence from the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 2019.

He is currently a Postdoctoral Researcher with Ben-Gurion University of the Negev.

Kamali Sripathi received the Ph.D degree in biophysics from the University of Michigan, Ann Arbor, MI, USA, in 2014.

She is currently a Postdoctoral Scholar with the Genome Center, University of California–Davis, Davis, CA, USA.



Hyunsoo G. Kim received the B.S. degrees in molecular and cell biology, and in psychology from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2015. She is currently working toward the Ph.D. degree in microbiology with the Department of Biomedical Engineering, University of California–Davis, Davis, CA, USA.

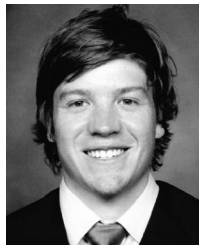
She is also affiliated with the Genome Center, University of California–Davis.

Marc T. Facciotti received the Ph.D. degree in biophysics from the University of California–Berkeley, Berkeley, CA, USA, in 2002.

He is currently an Associate Professor with the Genome Center and the Department of Biomedical Engineering, University of California–Davis, Davis, CA.

Michele Igo received the Ph.D. degree in biology from Harvard University, Cambridge, MA, USA, in 1986.

She is currently Associate Dean of Undergraduate Academic Programs with the College of Biological Sciences, and a Professor with the Department of Microbiology and Molecular Genetics with the University of California–Davis, Davis, CA, USA.



Nicholas Hoernle received the B.Eng. degree in electronics and computer engineering from the University of Cape Town, Cape Town, South Africa, in 2014, and the M.E. degree in computational science and engineering from Harvard University, Cambridge, MA, USA, in 2018. He is currently working toward the Ph.D. degree in artificial intelligence with the School of Informatics, University of Edinburgh, Edinburgh, U.K.

David Karger received the Ph.D. degree in computer science from Stanford University, Stanford, CA, USA, in 1994.

He is currently a Professor of Computer Science, and a Member of the Computer Science and Artificial Intelligence Laboratory with the Massachusetts Institute of Technology, Cambridge, MA, USA.