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

Shay A. Geller, Nicholas Hoernle, Kobi Gal, Avi Segal, Amy X. Zhang, Hyunsoo G. Kim, David Karger, Marc T. Facciotti, Michele Igo

***Abstract*—Students' confusion can be a barrier for learning, contributing to loss of motivation and to disengagement with course materials. However, detecting students' confusion in large-scale courses is both time consuming and resource intensive. This paper provides three new approaches for confusion detection in online forums that combines input from both students and teachers. The first approach describes a labelling tree that facilitates manual labelling of posts exhibiting confusion by the course staff and significantly improves the interrliability measure between labellers. The second method classifies confusion based on rules that were inferred from students' hashtags that are used to convey emotions. The third approach uses machine learning models to detect confusion in students' posts. We demonstrate how these three approaches can be applied in an online forum of a large scale Biology course. We combine the benefits of the three approaches by 1) selecting posts for training the machine learning models based on the rules that were inferred from students' use of hashtags in their posts; 2) using the labelling tree to generate labels for evaluating the models. We show that pretrained language model (BERT) that classify confusion in posts based on raw text, was able to outperform traditional machine learning models that rely on feature engineering. This model was also able to generalize across different terms of the same course. This work empowers teachers with better technologies for detecting and alleviating confusion in online discussion forums that combine input from both teachers and students.**

***Index Terms*—Educational technology, Unsupervised learning, Prediction methods.**

I. Introduction

BEHAVIORAL and emotional cues exhibited by students in the classroom communicate their level of interest, whether they are confused, frustrated or excited by the material the instructor is presenting [1]. Observing students' emotional cues give faculty important insight into how students are thinking about the course content which they can use to support students and facilitate course design.

Developments in technologies and increasing course sizes have moved traditional forms of course content (such as textbooks, lecture notes, and research articles) online, requiring students to master fundamental concepts outside the classroom. This makes it much harder for faculty to observe how the students are learning and to make assessments about how to intervene productively on the students' behalf.

In particular, negative emotions exhibited by students when they interact with course material, (such as confusion, frustration, and boredom) impede students' progress. When left unresolved, these can lead to low achievement, dropout, or academic dishonesty [2], [3]. However, when teachers are able to respond to students experiencing negative emotions in time, students may often return to an engaged state [2], [4]–[7].

This paper studies several approaches for detecting students' confusion when using collaborative annotated-based online forums, where students can anchor comments about course readings in the margins of the course text [8]. The context provided by such forums has been shown to improve the quality of the discussion and promotes student-to-student feedback (when compared to traditional forums) [9], [10].

Understanding confusion in annotated-based forums is a multifaceted question and students and teachers may have different perspectives about how posts can exhibit confusion. To this end we suggest three new approaches for identifying confusion in this setting [11]. The first approach facilitates manual labelling using experts (course teaching staff). We designed a labelling tree that guides the course teaching staff how to determine which posts exhibit confusion, and how. 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 labelling tree significantly improved inter-reliability agreement between experts when compared to asking them to label posts without using the tree.

The second approach classifies posts exhibiting confusion based on students' use of hashtags within the posts. Students can 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]. We show that a naive labelling rule that checks for the presence of the pre-defined #confused hashtag is a sufficient, but not necessary condition for inferring confusion in a student's post. This helps to motivate a new labelling rule, which considers additional hashtags that convey question and help-seeking behaviours, to also be indicative of confusion. We show that this second rule is more aligned with teachers' judgement of confusion in the students' posts.

The third approach uses machine learning classifiers for detecting confusion in students' posts. Leveraging the vast number of posts with students' self-reported hashtags, we

Manuscript received June 9, 2018; revised May XX, 2019; accepted XXX XX, 2019. Date of publication XXX XX, 2019; date of current version XXX XX, 2019. *(Corresponding author: Kobi Gal)*

Shay Geller, Avi Segal and Kobi Gal are affiliated with the Dept. of Software and Information Systems Engineering at Ben-Gurion University, Israel. Nicholas Hoernle and Kobi Gal are affiliated with the School of Informatics at the University of Edinburgh. Amy Zhang is affiliated with the University of Washington. David Karger is affiliated with MIT. Hyunsoo Kim, Marc Facciotti and Michele Igo are affiliated with the University of CA at Davis.

created a training dataset of posts that were labeled according to the rules that were inferred from students' hashtags. The test set consisted of posts that were labeled for confusion by the course experts using the labelling tree. We showed that a model that was pretrained on general text corpora and used only on the raw text of the post as input, was able to outperform traditional machine learning models that rely on feature engineering. Also this model was able to generalize between different terms of the course. This makes it a good candidate model to be used in practice.

Our work 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 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 machine learning models for detecting confusion, even for cases where hashtags are not available. Moreover, by directly involving teachers in the process of validating our rules, we were able to arrive at non-trivial conclusions for what constitutes confusion in online forums.

This paper significantly extends a prior conference publication [13] in several ways. First, we provide a richer, multi-class definition of confusion in students' posts that is informed by the education literature. Second, we design a structured way to collect labels from experts using the labelling tree. Third, we provide new machine learning models for detecting confusion that outperform those in the conference paper, and are able to generalize well across different instances of the same course.

II. Related Work

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

A. Confusion in Students' Learning

Perkins [14] and Piaget [15] each present a working definition for students' confusion. Perkins [14] defines students' confusion as the opposite of "understanding," which is a students' ability to grasp content or to retain and actively apply their knowledge. To extend this, Piaget [15] frames confusion 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.

Plaut [11] proposes the presence of multiple facets of confusion. Out of the four facets that Plaut [11] 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, and teachers' explanations to, and expectations from, the students. *Cause* relates to the underlying reasons for the confusion. Examples of cause can be unclear directions from teachers, students' misconceptions about the course content/goals, insufficient preparations (i.e., did not do homework), being "tuned out" (day-dreaming) in class, or reading material that is insufficient or too complicated for their level of understanding. We can think of types of confusion as an answer to the question, "*What is the student confused by?*" and causes of confusion to answer "*What is the reason for the student's confusion?*". These definitions relate to our work in Section VII-C, where we show that certain *types* and *causes* of confusion are more difficult than others to identify in forum posts.

Importantly, Plaut [11] 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 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 labelling tree to assist with labelling confusion consistently.

Lodge et al., [16] and Arguel et al., [17] present novel and unintuitive results that show 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 [18], productive failure [19], impasse-driven learning [20], cognitive disequilibrium [21], and discovery based environments [22], inject confusion to the learning process in different ways and test its effect.

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 [5]. Beck et al., [23] also proposed the concept of "wheel-spinning," where confused students fail in mastering a skill in a timely manner. This may lead to them never mastering the skill due to associated negative emotions which may lead to disengagement. Shute et al., [24] study the effect of persistent confusion in educational games and they suggest that targeted interventions can help students overcome this confusion. They, therefore, 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[1] dataset. The dataset contains $29,604$ annotated forum posts from 11 Stanford University public online courses. Nine expert colleagues labeled 6 emotions conveyed in posts, including confusion. They used a 7-point Likert scale to measure the level of confusion [25]–[28]. A different approach was to rely on a single expert, from the course's teaching staff, to identify the presence of confusion in forum posts [13]. The course staff member was asked "to what extent does the following post exhibit the student's confusion on the relevant reading material?," and used a 4-point Likert scale with 1 being "no confusion" and 4 being "very confused" to measure the level of confusion.

[1] https://datastage.stanford.edu/StanfordMoocPosts/

Yang et al., [2] used human labellers from MTurk to label thousands of students' posts in Algebra and Microeconomics courses. The crowed-sourced labellers were asked to judge the level of confusion exhibited in the posts on a 4-point Likert scale. We extend these works in providing a more nuanced definition of confusion, that is informed by pedagogical research. While the MTurk approach is highly scalable, it does not involve human experts.

Zhang et al., [29] used student-provided hashtags in Nota Bene to label curiosity and confusion. They assumed a one-to-one mapping from hashtags to affective states, and did not evaluate their approach with human experts.

C. Automated Labelling of Confusion in Forum Posts

Most work on machine learning models for confusion detection were based on the Stanford MOOCPost dataset [2], [25]–[28]. We list main approaches below. Agrawal et al., [25] 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., [26] also considered community-related features of the post like the number of reads and the number of up-votes of each post. They showed that the accuracy of the classifier monotonically increases with confusion level of the post (measured on a Likert scale of 1-7). Wei et al., [27] used convolution and recurrent neural networks for the confusion detection task.

Pre-trained, deep learning-based language models achieved state of the art results on several common prediction tasks [30]. Clavié and Gal [28] used Bidirectional Encoder Representations from Transformers (BERT) to improve prediction on confusion detection in the MOOCPost dataset. We show the efficacy of using BERT for confusion detection in the Nota Bene forum, showing that it can outperform models using hand-defined features used by [13].

III. Setting and Research Questions

Our empirical methodology uses data from the Nota Bene (NB) platform, an open-source collaborative online annotation tool[2]. The course content (PDF, HTML, video files) is uploaded to the NB website by instructors. Students annotate the content by highlighting a passage of the textbook (called "the marked text") and by typing into a text field that appears in the margin. These annotations may be used to comment on the content or to ask questions about what the students 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. NB is used by hundreds of university courses with more than $40,000$ registered student users.

Figure 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 that discusses the "TCA Cycle", a key metabolic pathway discussed extensively in introductory biology courses. 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 and can be made by highlighting a section of text and commenting. One can explore the annotations by a) looking through the text and selecting a highlighted section of interest, bringing up the corresponding annotation(s) on the right panel and/or by b) 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 which highlights areas in the document associated with many (dark yellow) and few (light yellow) annotations.

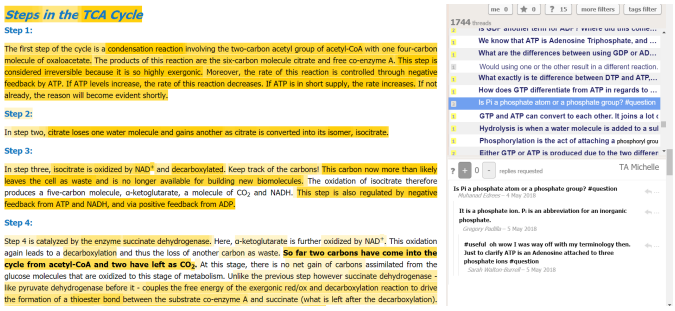

Fig. 1: Nota Bene GUI. Left: Course reading material. Highlighted text represents portions of text with an annotation. Right: A list of annotations.

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. 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 [9].

A. Bis2A course: Posts and Hashtags

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. Students in the course received reading assignments on material uploaded to NB. The students were required to provide "three meaningful posts" for each reading assignment in NB before each lecture. This encouraged active participation in forum discussions.

The NB interface allows students to generate pictorial representations (emojis) of students' emotions, using hashtags and text, (e.g. #confused, #useful, #frustrated) as shown in Figure 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

[2]http://nb.mit.edu/

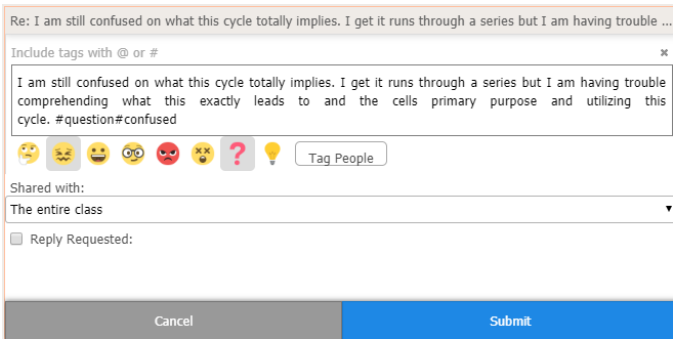

Fig. 2: Nota Bene hashtags GUI including post with the hashtags, and relevant emojis clicked (gray background surrounds them).

Hashtag	Count	Emoji	Hashtag	Count	Emoji
#confused	1,228		#help	420	
#curious	6,595		#question	7,574	
#interested	9,237		#useful	10,866	
#idea	9,673		#frustrated	65	

TABLE I: Hashtags, their associated emojis and the counts of their use in the summer term.

including at least one hashtag in at least one of their posts per lecture. Posts can be filtered by hashtags, allowing to navigate to posts that display specific emotions.

In total, out of $58,811$ posts is 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 in the course from using hashtags, supporting calls for providing students with opportunities for self-assessment [31].

Figure 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 3, 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 10, 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 media platforms [32]. Figure 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 4 or more of the 8 available 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%) while those that used all of the hashtags belonged to a small minority as well (2.2%).

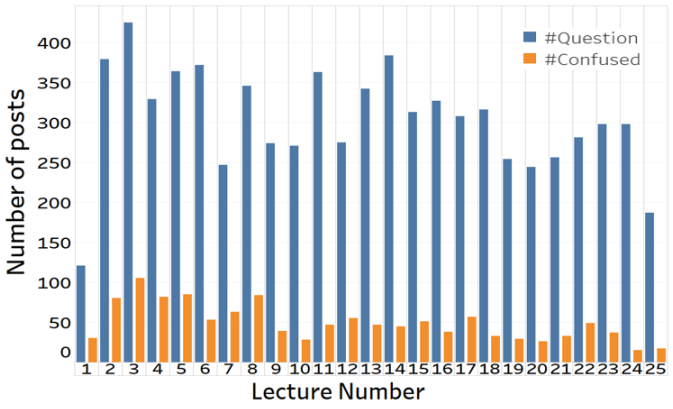

Fig. 3: Distribution of #confused and #question hashtags over lectures in the summer term

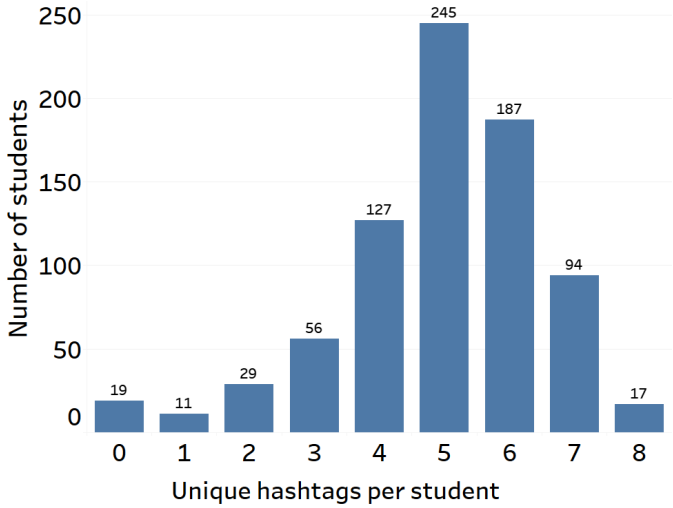

Fig. 4: Histogram of unique hashtag use by students in the summer term

B. Research Questions

Providing teachers with the possibility to use information about students' confusion can guide their interactions with the students and inform better course design. We propose that the students' self-reported affective states in the form of hashtags may be used to identify when, and on what content, the students are expressing confusion. Moreover, when hashtags are not present, as is 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 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 manually labelling of the posts. For example, the Stanford MOOCPosts dataset contains thousands

of posts that were manually labeled on a Likert scale of confusion by nine colleagues [33]. However, we identify that the notion of confusion is not universally agreed upon by experts (Section IV-A). To introduce rigour into the labelling process, we designed a labelling tree that leads experts toward more consistent labels.

Such labels can be used to train machine learning classifiers for determining confused posts based on linguistic cues and the vocabulary that is used in the students' posts [25], [26]. However, this approach imposes significant burden on the course staff, who in many cases are required to do the hand-labelling of the posts.

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

Research Question RQ1: Does the introduction of a structured labelling tree for guiding experts result in higher agreement, compared to using a Likert scale measure?

Research Question RQ2: Do students' hashtags agree with experts' judgements about what constitutes confused posts?

Research Question 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?

Sections IV and V describe the methodology and results for RQ1 and RQ2. Similarly, Sections VI and VII describe the methodology and results for RQ3.

IV. Methodology: RQ1 and RQ2

In this section we describe the methodology for addressing research questions 1 and 2. The methodology consists of two steps:

1) Design the labelling tree for collecting expert labels for confusion in Nota Bene (Section IV-A).
2) Using students' self-reported hashtags to construct rules for selecting confused posts (Section IV-B).

A. Labelling Tree for Experts

Previous works [13], [25] used a Likert scale to obtain confusion labels from experts. We consulted two experts from the Bis2A course staff to label over 200 posts using a Likert scale, but achieved low agreement ratios between the two experts (Kappa measure of 0.39 over 200 posts). This identified the need for a more structured approach to labelling confusion that reflects the taxonomy of different types of confusion from the pedagogical literature [11]. To this end, we developed a method that narrows the broad range of the concept of confusion by using a labelling tree, presented in Figure 5. Other works have demonstrated the benefits if using labelling trees to facilitate analysis of students' forum behavior [10], [34].

The labelling tree 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 that is 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: Declaring contradictions in the text, feeling lost about course material, making a wrong conclusion, and seeking information about course material or logistics. We also distinguished questions about material the student is supposed to know, 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.

B. Labelling Rules using Self-Reported Hashtags

To annotate posts for confusion using students' self-reported hashtags, a natural labelling 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 differnces [*sic*] between hydrogen bonds and ionic bonds were discusssed [*sic*]. I'm still having difficulty understanding the differences. #confused.

However, it is the case that some confused posts 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.

Moreover, the use of the #confused hashtag makes up only 3% and 11% ($1,228$ and $4,178$ posts in total) 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 super-set of those posts that are labeled with a #confused hashtag or that (2) the amount of confusion displayed in this dataset is significantly lower than that which has been noted by previous work [26], [29]. The above examples clearly show that the naive labelling rule, that uses only #confused, is an inadequate rule for detecting all of the confused posts and thus we are led to conclude that the former of these two cases is the correct one. We, therefore, need a new rule that describes confused posts which goes beyond #confused hashtag.

To find such a labelling 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. [35] that uses word embeddings to create vector representations for Twitter emojis. Similar to Wei et al., [27] that trained word embedding model on the posts from the Stanford MOOCPost dataset, we trained word embeddings using the Nota Bene posts. We then calculated the cosine similarity between the embedding vector of the #confused and the other hashtags to define a new labelling rule that considers other hashtags than just #confused.

To compute the embeddings, we trained a Skip-gram version of the Word2Vec model [36], with 50 hidden neurons in the embedding layer, on the text from all of the posts on the

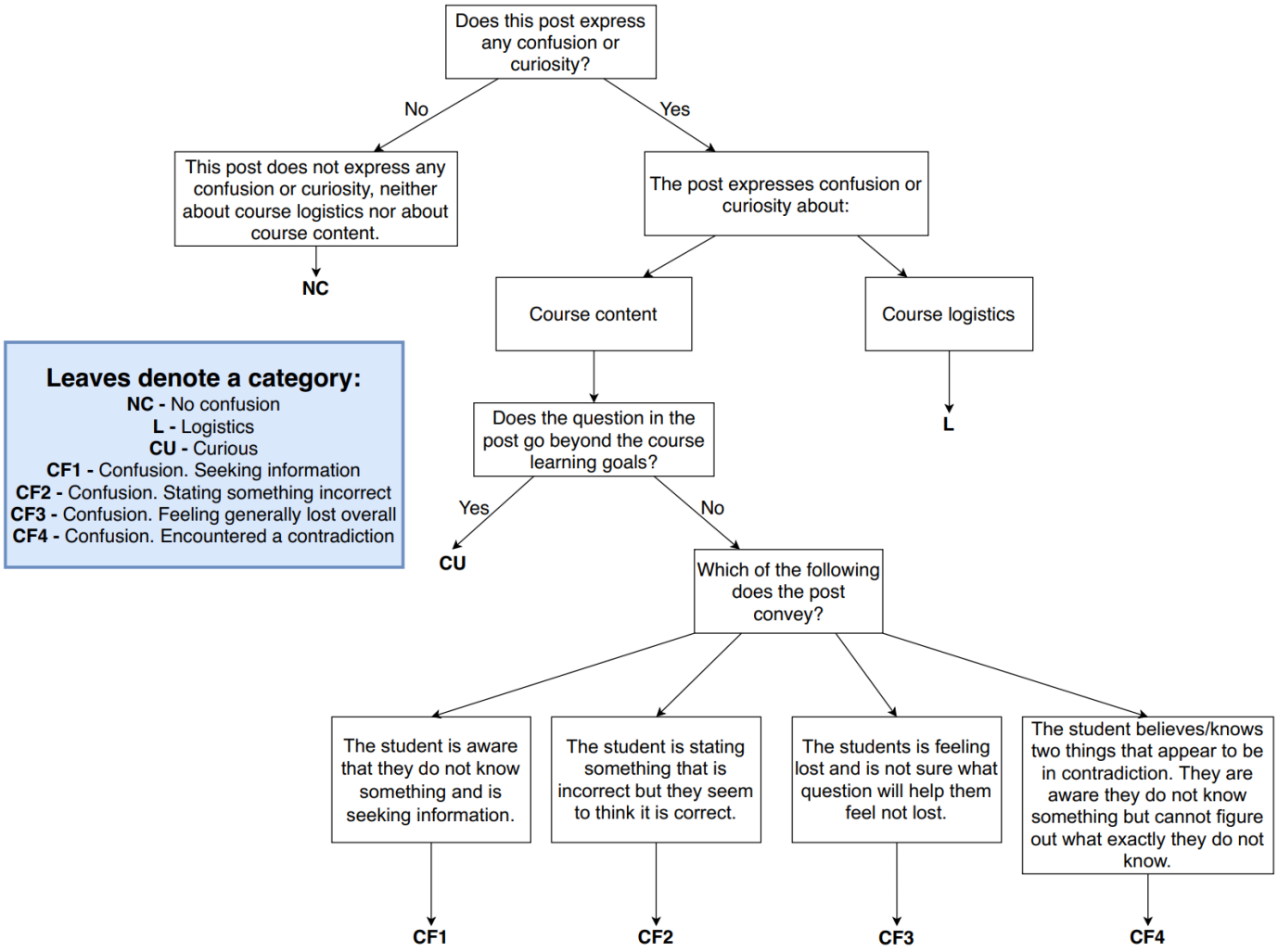

Fig. 5: Labelling tree for confusion

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.

Thus we suggest that new labelling rule should consider, beyond hashtag #confused, the other two #help and #question hashtags. Although the #curious hashtag also has a high cosine similarity to the #confused hashtag, we chose not to include it in the new labelling rule because Zhang et al., [29] were able to separate between these two hashtags with a ML classifier, hence they are used in different contexts.

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 (appears in less than 0.2% of all posts). Thus, we suggest the following two rules for detecting confused posts, the naive labelling rule and our data-informed rule:

$R1$**:** 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.

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 $R1$.

V. Results: RQ1 and RQ2

We begin by describing inter-labeler agreement, among experts, when they used the labelling tree from Section IV-A. Then we describe the agreement between the two labelling rules $R1$ and $R2$ and the experts.

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 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."

TABLE II: Definition and example for each leaf label of the labelling tree

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

TABLE III: Cosine similarity between the embedding for the #confused hashtag and the other hashtag embeddings (the closer the value is to 1, the more similar it is to #confused).

A. *Inter-labeler Agreement when using Labelling Tree*

To evaluate the labelling tree, we randomly picked 300 posts with no hashtags from each term of the Bis2A course, making 600 posts in total. We denote such posts as raw posts.

Three experts from the course staff labeled the 300 posts from the summer course term independently using the labelling tree of Figure 5. They achieved an intra-class 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 labelling tree to classify the confusion in the posts and achieved a Kappa score of 0.90.[3]

[3]We use the Kappa score because we are comparing two experts.

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

B. *Agreement Between Students' Self-Reported Hashtags and Experts*

To test the agreement between expert labels and our two labelling rules that are based on students' self-reported hashtags, we formed three groups of posts from the set of posts in the summer terms of Bis2A. Each group contained 50 posts with every post containing at least one hashtag. The first group includes 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 neither satisfy $R1$ nor $R2$.

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). Figure 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: group $R1$, group $R2 \setminus R1$, and 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

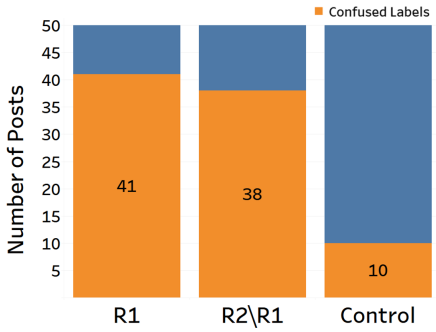

Fig. 6: Agreement between expert and posts in groups $R1$, $R2 \setminus R1$ and control group

expert and the posts in both groups $R1$ and $R2 \setminus R1$ 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 $R2 \setminus R1$, but this difference was not significant. The Kappa agreement between the expert and the posts in the $R1$ group was 0.28, and between the expert and posts in the $R2$ group was 0.55, which supports the evidence that the expert agrees more with posts when labeled using $R2$ 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 the rule $R2$ will improve the detection of confusion in this dataset. Thus, we have answered Research Question 2 in the affirmative.

VI. Methodology: RQ3

In this section we address RQ3, can we predict students' confusion in raw posts (that do not contain hashtags) using machine learning.

We first discuss in detail the datasets that are used for training and testing the different ML classifiers. We then introduce the feature engineering for pre-processing the data before classification. Finally, we present the machine learning models that are tested.

A. 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 were 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 "non-confused".

For the summer term, where we used three experts, the confusion label for each post was the majority label of the experts. 64 were labeled as "confused" and 236 posts labeled as "non-confused". 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, 52 "confused" and 236 "non-confused".

B. Feature Design

We applied standard pre-processing to the raw text. The preprocessing steps included (1) the removal of words that appeared less than 5 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 [37] were used to extract features from the raw text and the context of a post. The context of a post from NB consists of 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 which contains the highlighted-context). We used the following feature families:

1) General purpose textual features - Bigrams: We extracted word-level bigrams, which are general-purpose textual features [38], [39], from the Post-context. The extracted word-level bigrams' tokens counts were normalized using TF-IDF scores for each token [40].

2) Linguistic features: Linguistic Inquiry and Word Count (LIWC) [41] is a software package that analyzes text and extracts multiple categories of features that represent emotions, attention, sentiment, etc. This tool was widely used in similar studies of affective states [2], [42], [43].

For all of the contexts, we used the following features[4]: "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.

Many confused posts are located closely to paragraphs with a figure/plot or some example or explanation. We, therefore, further extract features over the Paragraph-context that can help us identify such paragraphs, like whether the paragraph contains the words "figure," "example',' "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, 100m).

Similar to [2], we also tried to identify sentences that explicitly express some confusion. To achieve 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."

3) Sentence complexity: Prior work on text readability have used features that capture the complexity of sentences [44]. Complex sentences could indicate difficult paragraphs, which may lead to more confused posts. 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.

[4]LIWC features description can be found here: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

4) Direction and Action items: The readings contains descriptors that guide understanding towards 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.

5) Question features: Questions are commonly used to express confusion [2]. To identify the questions, we extracted the following features over the Post-context: the number of 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).

C. Models

We experimented with several common machine learning classifiers. The training set for each model consisted of all posts with hashtags in 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 (SVM), decision trees, random forests and XGBOOST. Logistic regression with L1 regularization consistently outperformed the other models using 10-fold cross validation.

We compared the performance of the logistic regression model, with the custom features defined in section VI-B, with the Bidirectional Encoder Representations from Transformers (BERT) model, a pre-trained, transformer based language model that is commonly used in state-of-the-art natural language tasks [30]. This model is pre-trained on huge general text corpus from books and Wikipedia articles but was fine-tuned on the Bis2A posts. We fine-tuned the BERT model with two additional learning epochs, with a maximal sentence length of 512 words, using the *bert-base-uncased* configuration[5].

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

VII. Results: 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 hyper-tuned using 10-fold 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) AUC. Davis and Goadrich [45] show that the ROC curves can present over-optimistic results with skewed data.

[5]https://huggingface.co/transformers/model_doc/bert.html

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

TABLE IV: Within term results. Models trained and tested on summer term (top) and winter term (bottom).

Thus, it is preferable to use Precision-Recall (PR) curves as a reliable alternative [46]–[48]. As the model becomes both more correct and more confident about its predictions, the higher the PR-AUC number becomes, trending towards 1 for a perfect classifier.

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

- LR+B: A logistic regression model that is only trained on the word level bigrams features;
- LR+B+CW: The LR+B classifier with inverse class-weight re-balancing applied;
- LR+B+F: A logistic regression model that is trained on the word-level bigrams and the domain specific features;
- LR+B+F+CW: The LR+B+F classifier with inverse class-weight re-balancing applied;
- BERT: A BERT model that is just fined-tuned directly on the posts' raw text;
- BERT+D: The BERT model with down-sampling re-balancing applied.

A. Classification of Confusion Within Terms

Table IV compares the classifier performance on the summer (top) and winter (bottom) expert-labeled datasets respectively.

In both terms, the BERT models outperform the other models in all metrics except accuracy. This is also true for the winter term, where the BERT model 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. Furthermore, even though some of the logistic regression models, like LR+B+CW, achieve competitive results to the BERT models in the F1 metric, the PR-AUC metric shows that the BERT models are much more confident in their predictions.

Moreover, the results show that class re-balancing techniques on the training data result in an increase in recall and F1 across all models but a decrease in precision.

Finally, adding the domain-specific features did not contribute to better model performance on the summer term. Notably, the LR+B+CW model out-performs LR+B+F+CW in every metric. This is not consistent with the results from

Summer

Model	Precision	Recall	F1	Accuracy	PR-AUC
LR+B	$\downarrow 0.05$	$\downarrow 0.09$	$\downarrow 0.09$	$\downarrow 0.03$	$\downarrow 0.05$
LR+B+CW	$\downarrow 0.06$	$\downarrow 0.02$	$\downarrow 0.04$	$\downarrow 0.02$	$\downarrow 0.04$
LR+B+F	$\downarrow 0.05$	$\downarrow 0.09$	$\downarrow 0.09$	$\downarrow 0.03$	-
LR+B+F+CW	$\downarrow 0.03$	$\uparrow 0.04$	$\uparrow 0.02$	$\uparrow 0.01$	$\downarrow 0.01$
BERT	$\downarrow 0.04$	-	$\downarrow 0.01$	$\downarrow 0.01$	$\downarrow 0.03$
BERT+D	$\uparrow 0.04$	$\uparrow 0.03$	$\uparrow 0.03$	$\uparrow 0.02$	$\uparrow 0.01$

Winter

Model	Precision	Recall	F1	Accuracy	PR-AUC
LR+B	$\uparrow 0.05$	$\uparrow 0.19$	$\uparrow 0.14$	$\uparrow 0.03$	$\uparrow 0.05$
LR+B+CW	$\uparrow 0.09$	$\downarrow 0.02$	$\uparrow 0.04$	$\uparrow 0.03$	$\uparrow 0.05$
LR+B+F	$\uparrow 0.02$	$\uparrow 0.05$	$\uparrow 0.04$	$\uparrow 0.01$	$\uparrow 0.06$
LR+B+F+CW	$\uparrow 0.08$	$\uparrow 0.02$	$\uparrow 0.05$	$\uparrow 0.02$	$\uparrow 0.03$
BERT	$\uparrow 0.05$	$\downarrow 0.04$	-	-	$\uparrow 0.04$
BERT+D	$\uparrow 0.02$	-	$\uparrow 0.01$	-	$\uparrow 0.05$

TABLE V: Between terms results. Models trained and tested on different semesters. Results on summer term (top) and winter term (bottom). $\uparrow, \downarrow$ - represent an increase or decrease over same term's results from Table IV.

the winter term, where the domain features did lead to an increase in model performance across all metrics. This shows that a complex relationship exists between the contributions of the domain specific features across different terms.

B. 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 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 and tested on the summer dataset, in that 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 vs. 35,000 posts). Despite this fall in performance, the decrease in all model's performance is not more than 0.09 on the F1 measure. The results show some evidence that the classifiers generalize across the contexts of terms.

C. 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 that were collected in the within-term setting. Table VI presents results of the number of mis-classified posts for summer and winter terms.

The table shows that BERT-D performed best on the two classes with the most labels, achieving almost perfect prediction performance for the "No Confusion" (NC) and "Confusion - Seeking Info" ($CF1$)) labels.

	Summer			Winter		
	Total	Mis-classified	%	Total	Mis-classified	%
CF1	51	3	0.06	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	0.02	219	1	0.004
L	2	2	1	3	3	1
CU	13	11	0.85	14	14	1

TABLE VI: Error analysis of Bert+D model over expert-labeled posts on the summer and winter terms.

However, the BERT-D model performs poorly on the other labels, namely "Confusion - Wrong" ($CF2$), "Confusion - Lost" ($CF3$), and "Confusion - Contradiction" ($CF4$), and "Curious" (CU) that occur infrequently in the datasets. Although the sample size is low, we see a trend that the mis-classification rate on these posts is high, with a vast majority of these posts getting mis-classified 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. As the 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 labelling, goes beyond the ability of the natural language techniques that we applied. It is also important to note that the only labelling system to correctly identify these posts is that which depends on the expert labellers. Even the $R2$ labelling 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 which go beyond the content of the course. As with the $CF2$ label, we can neither expect the students nor the classifier to identify this label as it requires a thorough understanding of the Biology course curriculum.

VIII. Discussion

The results of our three research questions highlight the difficulty that arises in labelling 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 VII-C) shows there exist examples where this is not the case. Second, teachers and domain experts also represent a source for labelling 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. Labelling Confusion in Forum Posts

The results of the three research questions presented in this work reflect the different perspectives about students' confusion [11] as it is exhibited in online forums. We found

that using the labelling 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 labelling. 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, (2) are not always confident about the bounds of the scope of the course.

While we motivate the practical use of the labelling tree, we emphasize that was not meant to provide 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 while 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 labelling 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' cannot identify 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 $\frac{1}{3}$ 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 then that relying solely on students self-reported hashtags is insufficient for labelling 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 that used the labelling 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) make 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 entirely expected. The two reasons for the poor performance on these labels are (1) the classifiers were trained on the data from students self-reporting which are also 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 machine learning models for this task.

Our work highlights the importance of building machine learning models for detecting confusion that are informed by the pedagogical literature. The separation to multiple types of confusion entailed in the tree structure is an important expressive power that lacks in other labelling methods that only rely on Likert-scale methods. Therefore, we conclude that given a student's post, an expert course staff member who uses the labelling 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, are best identifiable by experts. Despite this, we show that the students' use of self-reported hashtags were very useful in training the Machine Learning models.

Finally, the course staff stated that labelling 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). This observation led them to consider adding clarifications or examples to specific portions of the text to help resolve misconceptions. The labelling tree, combined with students' posts in the discussion forums, can also be used as a pedagogical tool for teachers to obtain insights and resolve issues in the reading material.

IX. Conclusion and Future Work

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

The first research question discuss methods for labelling confusion by experts. We show that confusion is prone to multiple interpretations, even by course staff members, and simply using Likert methods for labelling confusion can lead to poor agreement between the experts. We propose a labelling tree that composed 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 that is exhibited by the post. This tree shown to produce consistent labels between experts by facilitating the labelling process in terms of coherence.

The second research question asks whether students' self-reported hashtags agrees with experts' judgement about what constitutes 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 labelling rules, the naive one that uses only #confused to signify a confused post, and the similarity 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 pre-trained language models (BERT) that accepts only the raw text of the post as an input. We show that classifiers trained hashtag-labelled posts perform well on expert-labelled 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 that shows that some types of confusion are harder to detect. This encourages the importance of the hashtags as valuable training data, and the fine-grained confusion labels obtained when experts used the labelling tree as a way to understand the models' true blind spots. In future work, we wish to extend the machine learning models to improve prediction of the less popular confusion types such as CF2 (wrong statement is not aware of). Also, we wish to design a confusion "heat-map" 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 heat-map can guide teachers in their design of the course.

REFERENCES

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

[2] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rose, "Exploring the effect of confusion in discussion forums of massive open online courses," in *Proceedings of the second (2015) ACM conference on learning@ scale*. ACM, 2015, pp. 121–130.

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

[4] B. Grawemeyer, M. Mavrikis, W. Holmes, and S. Gutierrez-Santos, "Adapting feedback types according to students' affective states," in *International Conference on Artificial Intelligence in Education*. Springer, 2015, pp. 586–590.

[5] S. D Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.

[6] M. M. T. Rodrigo, R. S. Baker, and J. Q. Nabos, "The relationships between sequences of affective states and learner achievement," in *Proceedings of the 18th International Conference on Computers in Education*, 2010, pp. 56–60.

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

[8] S. Zyto, D. Karger, M. Ackerman, and S. Mahajan, "Successful classroom deployment of a social document annotation system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1883–1892.

[9] K. Miller, S. Zyto, D. Karger, J. Yoo, and E. Mazur, "Analysis of student engagement in an online annotation system in the context of a flipped introductory physics class," *Physical Review Physics Education Research*, vol. 12, no. 2, p. 020143, 2016.

[10] E. Yogev, K. Gal, D. Karger, M. T. Facciotti, and M. Igo, "Classifying and visualizing students' cognitive engagement in course readings," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, 2018, p. 52.

[11] 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.

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

[13] S. A. Geller, N. Hoernle, K. Gal, A. Segal, A. X. Zhang, D. Karger, M. T. Facciotti, and M. Igo, "# confused and beyond: detecting confusion in course forums using students' hashtags," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, pp. 589–594.

[14] D. N. Perkins, *Smart schools: Better thinking and learning for every child*, 1992, no. Sirsi) i9780028740188.

[15] J. Piaget, *The psychology of intelligence*. Routledge, 2005.

[16] J. M. Lodge, G. Kennedy, L. Lockyer, A. Arguel, and M. Pachman, "Understanding difficulties and resulting confusion in learning: An integrative review," in *Frontiers in Education*, vol. 3. Frontiers, 2018, p. 49.

[17] A. Arguel, L. Lockyer, O. V. Lipp, J. M. Lodge, and G. Kennedy, "Inside out: detecting learners' confusion to improve interactive digital learning environments," *Journal of Educational Computing Research*, vol. 55, no. 4, pp. 526–551, 2017.

[18] E. L. Bjork, R. A. Bjork *et al.*, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," *Psychology and the real world: Essays illustrating fundamental contributions to society*, vol. 2, no. 59-68, 2011.

[19] M. Kapur, "Productive failure," *Cognition and instruction*, vol. 26, no. 3, pp. 379–424, 2008.

[20] K. VanLehn, "Toward a theory of impasse-driven learning," in *Learning issues for intelligent tutoring systems*. Springer, 1988, pp. 19–41.

[21] A. C. Graesser, S. Lu, B. A. Olde, E. Cooper-Pye, and S. Whitten, "Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down," *Memory & cognition*, vol. 33, no. 7, pp. 1235–1247, 2005.

[22] L. Alfieri, P. J. Brooks, N. J. Aldrich, and H. R. Tenenbaum, "Does discovery-based instruction enhance learning?" *Journal of educational psychology*, vol. 103, no. 1, p. 1, 2011.

[23] J. E. Beck and Y. Gong, "Wheel-spinning: Students who fail to master a skill," in *International conference on artificial intelligence in education*. Springer, 2013, pp. 431–440.

[24] V. J. Shute, S. D'Mello, R. Baker, K. Cho, N. Bosch, J. Ocumpaugh, M. Ventura, and V. Almeda, "Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game," *Computers & Education*, vol. 86, pp. 224–235, 2015.

[25] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke, "Youedu: addressing confusion in mooc discussion forums by recommending instructional video clips," 2015.

[26] Z. Zeng, S. Chaturvedi, and S. Bhat, "Learner affect through the looking glass: Characterization and detection of confusion in online courses." in *EDM*, 2017.

[27] X. Wei, H. Lin, L. Yang, and Y. Yu, "A convolution-lstm-based deep neural network for cross-domain mooc forum post classification," *Information*, vol. 8, no. 3, p. 92, 2017.

[28] B. Clavié and K. Gal, "Edubert: Pretrained deep language models for learning analytics," *arXiv preprint arXiv:1912.00690*, 2019.

[29] A. X. Zhang, M. Igo, M. Facciotti, and D. Karger, "Using student annotated hashtags and emojis to collect nuanced affective states," in *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. ACM, 2017, pp. 319–322.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] J. A. Ross, "The reliability, validity, and utility of self-assessment," 2006.

[32] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, 2010, pp. 241–249.

[33] A. Agrawal and A. Paepcke, "The stanford moocposts dataset," 2014.

[34] X. Wang, M. Wen, and C. P. Rosé, "Towards triggering higher-order thinking behaviors in moocs," in *Proceedings of the Sixth International Conference on Learning Analytics Knowledge*, ser. LAK '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 398–407. [Online]. Available: https://doi.org/10.1145/2883851.2883964

[35] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, "emoji2vec: Learning emoji representations from their description," *arXiv preprint arXiv:1609.08359*, 2016.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[37] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[38] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242. Piscataway, NJ, 2003, pp. 133–142.

[39] L. S. Jensen and T. Martinez, "Improving text classification by using conceptual and contextual features," Ph.D. dissertation, Citeseer, 2000.

[40] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[41] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[42] O. Almatrafi, A. Johri, and H. Rangwala, "Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums," *Computers & Education*, vol. 118, pp. 1–9, 2018.

[43] T. Atapattu, K. Falkner, M. Thilakaratne, L. Sivaneasharajah, and R. Jayashanka, "An identification of learners' confusion through language and discourse analysis," *arXiv preprint arXiv:1903.03286*, 2019.

[44] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 186–195.

[45] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[46] M. Craven and J. Bockhorst, "Markov networks for detecting overalpping elements in sequence data," in *Advances in Neural Information Processing Systems*, 2005, pp. 193–200.

[47] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial intelligence in medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[48] S. Kok and P. Domingos, "Learning the structure of markov logic networks," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 441–448.

Kobi Gal is an Associate Professor at the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev, and Reader at the School of Informatics at the University of Edinburgh. He received the PhD degree from Harvard University in 2006.

Avi Segal is a postdoctoral researcher at the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev.

Shay A. Geller is a master's student at the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev and is supervised by Kobi Gal.

Nick Hoernle is a PhD student at the University of Edinburgh and is supervised by Kobi Gal.

Michele Igo is Associate Dean in the College of Biological Sciences and Professor in the Department of Microbiology and Molecular Genetics at the University of California, Davis.

Marc T. Facciotti is an Associate Professor in the Genome Center and the Department of Biomedical Engineering at the University of California, Davis.

Hyunsoo G. Kim is a PhD candidate advised by Marc T. Facciotti in the Genome Center and Department of Biomedical Engineering at the University of California, Davis.

Amy X. Zhang is an Assistant Professor in the Allen School of Computer Science and Engineering at the University of Washington.

David Ron Karger is a professor of computer science and a member of the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.