5-2017

Towards Student Engagement Analytics: Applying Machine Learning to Student Posts in Online Lecture Videos

Nicholas R. Stepanek
University of Nebraska at Omaha

Towards Student Engagement Analytics: Applying Machine Learning to Student Posts in Online Lecture Videos

A Thesis

Presented to the
Department of Information Science and Technology
and the
Faculty of the Graduate College
University of Nebraska

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science

University of Nebraska at Omaha

by
Nicholas R. Stepanek
May 2017

Supervisory Committee
Brian Dorn
Yuliya Lierler
Robin Gandhi

TOWARDS STUDENT ENGAGEMENT ANALATYICS: APPLYING MACHINE LEARNING TO STUDENT POSTS IN ONLINE LECTURE VIDEOS

Nicholas R. Stepanek, MS

University of Nebraska, 2017

Advisor: Brian Dorn

The use of online learning environments in higher education is becoming ever more prevalent with the inception of MOOCs (Massive Open Online Courses) and the increase in online and flipped courses at universities. Although the online systems used to deliver course content make education more accessible, students often express frustration with the lack of assistance during online lecture videos. Instructors express concern that students are not engaging with the course material in online environments, and rely on affordances within these systems to figure out what students are doing. With many online learning environments storing log data about students usage of these systems, research into learning analytics, the measurement, collection, analysis, and reporting data about learning and their contexts, can help inform instructors about student learning in the online context.

This thesis aims to lay the groundwork for learning analytics that provide instructors high-level student engagement data in online learning environments. Recent research has shown that instructors using these systems are concerned about their lack of awareness about student engagement, and educational psychology has shown that engagement is necessary for student success. Specifically, this thesis explores the feasibility of applying machine learning to categorize student posts by their level of engagement. These engagement categories are derived from the ICAP framework, which categorizes overt student behaviors into four tiers of engagement: Interactive, Constructive, Active, and Passive. Contributions include showing what natural language features are most indicative of engagement, exploring whether this machine learning method can be generalized to many courses, and using previous research to develop mockups of what analytics using data from this machine learning method might look like.

Acknowledgements

The completion of this thesis would not have been possible without those who have supported me along the way. Thank you to my committee members, Dr. Brian Dorn, Dr. Yuliya Lierler, and Dr. Robin Gandhi. In particular, I would like to thank Brian Dorn for mentoring me these past two years and giving me the opportunity to work in the BRIDGE lab. Without him I would not be graduating with the amount of knowledge and experience I have gained while studying here. Thanks to Suzanne Dazo and Robert Gibbs who helped me with the coding of over 4700 posts written by students. Thanks to my family who have always supported me, and all of my friends who have helped keep me sane throughout my education. Finally, thanks to all of the other lab members who have helped me edit and review portions of this work and my thesis defense.

This work is funded in part by the National Science Foundation under grant IIS-1318345. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Using online avenues to deliver course content is becoming more popular in higher education, not only to be used in online classes, but also for providing lectures outside of class in flipped courses [3]. Additionally, MOOCs (Massive Open Online Courses) are becoming increasingly prevalent and aim to make higher education more accessible to those interested for little to no cost. Commonly, the systems used to deliver course content online store fine-grained, click-level data about student usage, leaving the door open for a myriad of in-depth analyses on student behaviors. Some examples include sentiment analysis of text artifacts [48], clustering activity patterns [29], language analysis [47], and social network analysis [42]. However, one problem with the storage and analysis of all these student logs is figuring out how to use the results for improvement in instruction.

One method of helping instructors receive the benefits of all of this log data is through research in *learning analytics*, the measurement, collection, analysis, and reporting of data about learning and their contexts [19]. These analytics can help us deliver high-level information to instructors as students use the system, enabling instructors to use the results to modify instruction in near real time. Learning analytics can deliver many different kinds

of information based on what data is collected in a system, but this thesis focuses on one aspect of learning, student engagement. Figuring out how student engagement manifests itself in student logs and what to do with that information is a major area of research with these online learning systems [1, 23, 26, 47, 53]. Additionally, information about how engaged students are with lecture videos is something instructors using an online video lecture system have commonly asked for during interviews [18].

The goal of this thesis is to address the need for more instructor awareness about student engagement during online learning. To do this, text artifacts produced by students in an asynchronous media platform that multiple universities use to host lecture videos are utilized. Using the ICAP framework, "a taxonomy that differentiates four modes or categories of engagement, based on the overt behaviors displayed or undertaken by students" [8], student texts are categorized based on their engagement with the course material and a machine learning algorithm is trained for future classification of student engagement. Specifically, the following research questions will be addressed:

- What language features are most important for classifying student posts by engagement?
- Can some classifier be generalized to work on any course, or is success dependent on training data from that specific course?
- To what extent can machine learning automatically categorize engagement in lecture videos using text artifacts as data?

The overarching objective is to test the feasibility of using text artifacts to communicate engagement information to instructors using machine learning, as opposed to instructors needing to manually analyze every artifact produced by students. The first question helps

inform the machine learning stage by discovering what features will lead to the best accuracy, in addition to providing a qualitative description of how engagement manifests itself in text artifacts. The second question helps discover what a classifier identifying engagement looks like when generalized to multiple courses. The last question addresses whether a machine learning algorithm to categorize engagement is possible, preceding the development of a learning analytic conveying engagement information to instructors.

The rest of this chapter further introduces the contexts of concern (MOOCs and flipped courses), why engagement is important to study, and how instructors play a crucial role in student engagement. Chapter 2 covers work related to this thesis and introduces the ICAP framework, a psychological framework for student engagement. Chapter 3 describes the methodology of gathering the dataset, coding annotations, and the analyses used on the data set. Chapter 4 contains the results of those analyses and discussions of the interesting results. Chapter 5 uses the results and discussions from the previous chapter and prior work with TrACE instructors to produce mock-ups of what an engagement analytic might look like. Chapter 6 summarizes and concludes this thesis.

1.1 MOOCs, Flipped Courses, and Engagement

One area that online lecture videos and online learning is becoming more prevalent in is in MOOCs, or Massive Open Online Courses. Through MOOCs, classes are delivered through the web using online lecture videos and use forms of automation such as automatically graded quizzes to reduce the workload of instructors. This allows instructors to virtually replicate the classroom experience for any number of students without requiring the same instructor to student ratio typically found in universities [25]. The lecture videos, reading

materials, quizzes, discussion boards, and more, are used to deliver all course content and replace the experience of lectures in a traditional classroom.

Many leading MOOC platforms such as EDx and Coursera allow anyone to take their classes for free [51]. This has produced extremely high enrollment numbers for some popular courses, with some classes having over 100,000 students initially enrolled [27]. However, this also means that instructors are not often capable of giving individual help and attention to students who may need it. The concept of office hours from traditional university classrooms simply does not exist for large MOOCs, and the extremely high dropout rates [27] suggests that students in MOOCs may find it difficult to engage with the material, other students, and instructors. The ability for MOOCs to scale to any number of students comes with the drawback of instructors not being able to pay much personal attention towards individual students and their progress. The impact that this lack of personal attention may be having will be discussed later in this section.

The second context in which lecture videos are seeing more use is in flipped classes. The flipped classroom model has been increasing in popularity in higher education [15, 41], which may be due to a gaining legitimacy of constructivist educational psychology and active learning in the minds of instructors [3]. In a flipped classroom, students independently consume traditional lecture material outside of class so that class time can be used for active learning, including problem and discussion based tasks [33]. The core of the flipped classroom is at the active learning component, and it is worth noting that the model only demands some form of learning outside of class to replace traditional lectures, not necessarily the use of online lecture videos. However, many instructors find them convenient in replacing the traditional lecture experience.

Another frequently touted benefit of the flipped course model is the ability for students

to revisit lecture material at their own pace [24]. If the content is overwhelming to a student at first, they can take a break, or go over that material again later. Recent empirical evidence however suggests that students do not often revisit lecture videos, let alone view the material for a first time before class when there are no participation grades associated with the videos [13]. A possible explanation may be the introduction of a problem similar to that of the MOOC. With online lecture videos, instructors are removed from that portion of the learning process, and students may become discouraged without the ability to ask questions and receive immediate feedback. Students have reported becoming discouraged during the lecture process because they feel like they have to learn the course material "all on their own" [40]. Instructors report that students in flipped classes express frustration due to the lack of assistance while watching videos and find students watching videos passively [39].

Discussed here have been various aspects of flipped courses and MOOCs that may affect student engagement, specifically a lack of instructor presence, but why is engagement so important to learning? Learning is the ongoing process of knowledge construction during interactions among learners, instructors, and resources [4]. Student engagement, or a student's level of interaction with other learners, instructors, and resources, is then vital to learning as these interactions resulting from engagement are where knowledge construction occurs. Many studies in online learning systems, discussed in the related works chapter, find correlations between engagement and student success or performance. The goal of this thesis is to provide a learning analytic that will inform instructors on student engagement so that they are more capable of instructor intervention, something instructors often lose when moving to online learning environments. To better understand how an instructor can affect engagement, the next section looks closer at how instructors can affect the learning

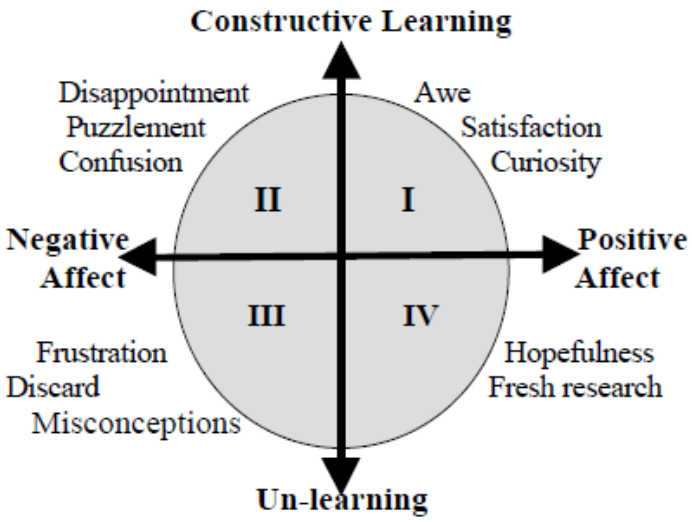

FIGURE 1.1: The affective states of learning, the x-axis referring to emotion and the y-axis referring to the construction or destruction of knowledge [31].

process.

1.2 The Affective States of Learning

To better understand the learning process and in what ways instructors affect student engagement, consider figure 1.1. Represented are the four affective states of learning [31] which are tied to emotions and type of learning that the student will experience during each state. For the rest of this paper, the states from figure 1.1 will be referred to as satisfaction (I), confusion (II), frustration (III), and hopefulness (IV) respectively. The four categories are split by two different attributes, the type of learning and emotional affect. Generally, students begin the learning process in the confusion or satisfaction state, and although different experiences can move the student anywhere, a typical learning process moves counter-clockwise around the circle [30].

Starting on the side of constructive learning, the goal state for students is the satisfaction state. In this state, the student is both experiencing constructive learning as well as a

positive emotional state, more specifically an emotional state where they are curious about the new material and confident that they are building an understanding of it. However, students who do not grasp as much of the course material upon initial explanation will start in the confusion state. In this state the student is still learning, but is experiencing a negative emotional state due to being dissatisfied about not immediately grasping all of the material. On the the other side, frustration occurs when a student is experiencing negative emotions and is unlearning, in that the student is gaining misconceptions about the course material. Kort [30] provides the example of a student writing a computer program, only to find that it does not work properly. The reason it does not work is due to a misconception the student holds, and upon realizing this the student experiences a negative emotional state. This is the most challenging state as the student needs motivation to move to the hopefulness state where misconceptions can be unlearned. The hopefulness state is both positive emotionally and a state of unlearning. In this state the student has acknowledged their misconceptions and is working to remove them, and will then likely move to confusion or satisfaction states when they are ready for constructive learning again.

1.2.1 Moving Students through the Affective States

Modeling learning in this way, it would seem obvious to try and ensure students are experiencing the satisfaction state as much as possible. However, confusion naturally occurs when students are learning something new and it is impossible to completely prevent misconceptions. Experiencing confusion is essential in learning as the affective state of confusion is positively correlated with learning outcomes [11]. Students who experience confusion during the learning process are more likely to learn more, or deeper, than students who learn in a comfortable environment that does not challenge their already established knowledge [17].

More intuitively, a student will not learn from material they already understand as they already know it. Additionally, students do not immediately understand new course content the first moment they are exposed to it, and it will most often take time working in the confusion and other states before the student can move to the satisfaction state. Knowing that students often move counter-clockwise through the affective states, it is crucial that students remain engaged enough and are provided enough assistance to move through phases of frustration. This creates a danger for learning in the online context, where instructors lack the control they typically have in traditional classrooms.

Instructors are the main mediators for ensuring students remain engaged and have their questions resolved, meaning that they play a vital role in ensuring that students make it to the satisfied state. In the traditional classroom, students who are confused or frustrated have the opportunity to ask questions for further explanations. In the online lecture video context, there is no method for students to quickly contact the instructor with questions (though many researchers are developing systems to replace these lost affordances, the system used in this thesis being one of them [16, 43, 52]). As discussed earlier, this leaves students with the feeling of needing to learn the material all on their own, and frustrated students then have no external aid in resolving their misconceptions as they would in a traditional classroom.

To summarize this portion, it is more likely that students become disengaged and therefore experience frustration in the online context. Providing instructors more information about student engagement in the online context would greatly improve their ability to quickly identify students stuck in the frustration state to then push their learning forward through the affective states. Alternatively, students may give up on learning entirely, possibly producing the high drop rates in MOOCs seen today [27] and lack of student viewing

in flipped courses [13].

1.2.2 Frustration and Learning Outcomes

Another concern with instructors having less ability to engage with students in the online context is that this may be disproportionately affecting weaker students. Looking closer to the roots of active learning, another way to look at the learning process is that the new material needs to be just outside the realm of the students already established knowledge for acquisition. Vygotsky coined this area as the zone of proximal development. He defined the zone of proximal development as the distance between a child's "actual developmental level as determined by independent problem solving" and the "potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" [49], and that all new knowledge is constructed in this zone. Generally, students who are attempting to learn course material that is further outside of their zone of proximal development will spend more time in the frustration state. On the other hand, students already familiar with more background knowledge and preparation for the course will experience less frustration and require less intervention from the instructor.

Students spending more time in the frustration is very troubling for learning outcomes, especially when weaker students who need the help are most affected. Two studies on students' affective states have found that frustration is often a normal occurrence, but students spending longer periods of time frustrated is correlated with reduced learning outcomes [34, 35]. Course material that is outside a student's zone of proximal development will quickly lead a student to frustration, and there is nothing the instructor can do in the online context excluding advanced collaborative systems to intervene. The negative emotional state may lead the weaker students to give up comprehending the material, significantly

impacting learning outcomes. A learning analytic that enables faster instructor intervention in the online context may enable instructors to give weaker students the attention needed to excel in the course.

1.3 Summary

To conclude this section, analyzing student engagement in online learning systems is an increasingly popular research topic, but systems for actually improving engagement are relatively few. Given the amount of log data stored in these kinds of systems, the challenge remains of finding methods to use with that data to improve student success. This thesis attempts to solve the problem of instructor awareness about student engagement during lecture videos in online systems, as student engagement is crucial to success and instructors lose the power of intervention they typically have in traditional classrooms. To accomplish this, machine learning methods will be applied to a collected data set of annotations to see if automatic categorization is possible under the ICAP framework. If so, learning analytic systems can be built providing this engagement data to instructors to enable intervention.

This thesis makes the following contributions:

- Features that indicate engagement - The language attributes that text artifacts in the courses studied that largely determine engagement are found through statistical analyses. Performance evaluation of a machine learning method utilizing these features shows that the machine learning method proposed in this thesis is possible.

- How engagement changes for different courses - Through statistical analyses comparing the different courses studied, it is shown that the features studied in this thesis

have the same relationships with engagement between courses, but more work is likely needed to generalize this machine learning method to many courses.

- Initial designs of engagement learning analytics - With a machine learning method that is able to categorize new text artifacts into the ICAP categories, the results from this thesis and prior research are used to develop mockups of what a learning analytic built on top of this machine learning method might look like.

Chapter 2

Related Work

This section is a discussion of recent literature that has analyzed MOOC behaviors through click-stream data, their findings, and the implications of them on this thesis. Next is a discussion of studies that have applied text analyses and natural language processing on MOOC and learning system data. Then, the ICAP framework will be introduced. A discussion on prior research with it in education and text artifact contexts will describe why the ICAP framework is fit for use in this thesis. Finally, a summary describes the implications of the discussed literature on this thesis.

2.1 Analyses of MOOC Behaviors

With large-scale MOOCs such as EdX and Coursera that collect click stream data about thousands of users, the challenge of analyzing all that data in meaningful ways for both instructors and researchers arises. One major area of research with this data is developing a model for MOOCs and learning management systems that predicts student performance, which may then be used to inform instructors about students who may need extra help succeeding in the course. Brinton and Chiang looked at predicting course performance in

two Coursera classes by analyzing variables generated by students video viewing behaviors [6]. The variables found to be correlated to performance were those indicating engagement with the videos, including skipping within the video, changing the playback speed, pausing the video, and rewinding at least once. Coleman et al. attempted to predict certification in an MITx course on EdX [10]. They considered video watching behaviors, assignment viewing behaviors, and quiz performance to create a model that could predict certification with 81% accuracy. The models developed in these two pieces focus on using click-stream from the MOOC, providing an objective look at what students are doing in the system, and MOOC studies looking at different issues generally rely on similar metrics derived from log data.

A study on the same MITx course in EdX sought to report correlations between certificate earners and more qualitative data collected through surveys, such as study strategies and educational background, as a precursor to their performance prediction model [5]. There were correlations between success and past degrees and experience in mathematics, and a strong correlation between people who reported collaborating with another student in the course or someone who has expertise in the field and success. Using demographic data as in this study paints a broader picture of each student, but requires extra participation on the part of the student. Other studies with learning management systems that already collect this kind of data have tried to combine both click-stream data and demographic information to identify students in need of help.

Wolff, Zedenek et al. created a predictive model to identify at-risk students using Open University [50]. By combining demographic data, assessment scores, and click-level data within Open University they were able to predict failing students with high confidence. Course Signals also predicts at-risk students, but allows students to know if they are labeled

at risk [2]. Signals was designed to give at-risk students extra help and alerts instructors about at-risk students to allow for prompt intervention. They concluded that Signals had a positive impact on overall grades, and a study on perceptions of Signals found that instructors saw in increase of students utilizing extra resources such as tutors and office hours [28].

The analyses in this section have looked at engagement through the lens of click-level behaviors and demographics in various systems with the goal of identifying performance. It is worth noting that all studies found correlations between student success and metrics that indicate engagement, such as video watching, building a stronger case for engagement being a worthy indicator to study in this thesis. Recently, researchers have began to apply NLP techniques to the text artifacts in these systems as another source of data to supplement the results from these strictly log and demographic data studies. Although log data provides objective insight into how students are using these systems, text artifacts go a step further by providing evidence into how students are actually thinking about the course materials.

2.2 Text Analyses in MOOCs

Although a much newer trend than analyses on click-level data, some researchers have began applying NLP methods to the text artifacts in online lecture systems, often with the same goal of predicting engagement and student performance. Wen et al. performed a sentiment analysis of all text documents produced by students in three different MOOC courses to see if sentiment was correlated with dropping out of the course [48]. The sentiments measured where simple positive/negative measurements based on the individual words in the texts. In a Python course, both positivity and negativity were correlated with dropping out, though in a fantasy writing course negativity was correlated with staying in the course. This seemed

due to users sharing stories often containing negative words as being more likely to stay in the course, leading to the conclusion that sentiment in text artifacts in very dependent on the context of the course. A course where students are commonly posting works of fiction to share with others will likely influence results.

Kavanović et al. developed a system to automatically classify text artifacts based on their level of cognitive presence in an online research course in software engineering [32]. Cognitive presence is one of the three constructs in the Community of Inquiry model for computer-mediated communication and is necessary for student success in online learning environments [21]. This framework describes cognitive presence as being the extent that participants are able to construct meaning through sustained communication. Kavanović et al. found some of the most important features to be word counts, text coherence metrics, the number of replies a message received, how deep a message was in a thread, similarity to nearby messages, and whether the message was first or last in a thread. They trained a random-forest classifier which achieved 70.3% accuracy in predicting cognitive presence. Random-forest classifiers are able to identify which features were most important for classification, and they provide a table of feature importance for research in similar contexts.

Crossley et al. noticing the lack of student success models combining both click-stream data and NLP analyses, developed a model for predicting student success in a MOOC based on both activity data and text artifacts [12]. The activity metrics they used included the average amount of times they accessed data in the MOOC per week they were active, the percentage of videos they had watched by their due date, how many times they accessed a forum, created a post, or commented, viewed a page in the course, and their number of assignment submissions. From these metrics, and a number of NLP metrics calculated from student posts in the discussion board, they applied a MANOVA to determine which ones

were most effective in identifying students who completed the course. Their most useful NLP features included high school essay score, total number of words produced, average post length, concreteness, tri-gram frequency, and semantic similarity between paragraphs.

These studies including text analyses help inform what features to use for classification in this thesis. Furthermore, the features that were found to be most useful for identifying cognitive presence and student success were again features describing student engagement. Next, the ICAP framework is introduced, used in this thesis as the theoretical framework for categorizing engagement.

2.3 ICAP

The ICAP framework attempts to classify the degree to which a student is engaged through their overt behaviors [8]. ICAP defines four modes of overt engagement behaviors: *interactive*, *constructive*, *active*, and *passive*. The framework specifies *overt* behaviors as it only classifies what is visible to others, or in the scope of this thesis text documents. It does not attempt to account for other thoughts of the student, only the product, though research discussed in this section has shown that the products will reflect the actual thought and engagement of the student.

The first category, *passive*, includes behaviors where students are not interacting with the course material in any overt way. This includes situations where students are only present for observing course material, like simply listening to a lecture or video. The *active* category means the student is physically manipulating the course material, by taking notes or highlighting a text, but is not inferring any new knowledge that goes beyond what is explicitly stated by the materials. *Constructive* behaviors are those where students build knowledge by comparing the materials with prior knowledge or by inferring new conclusions

with the course material. This is observed through actions such as self-explaining, generalizing information, or generating new ideas that go beyond the course content. The final and most highly engaged category, *interactive*, involves dialogue between two students. These behaviors construct knowledge by discussing different ideas and perspectives with peers that go beyond the explicit ideas in the course materials. For more examples of what types of behaviors would be classified as which of the modes of engagement, see 2.1. More succinct definitions of the four categories by Chi and Wylie follow:

Interactive - Dialogues where both partner's utterances must be primarily constructive with a sufficient degree of turn taking.

Constructive - Behaviors in which learners generate or produce additional externalized outputs or products beyond what was provided in the learning materials.

Active - Behaviors in which some form of overt motoric action or physical manipulation is undertaken.

Passive - Behaviors in which learners are oriented towards and receiving information from the instructional materials without overtly doing anything else.

	PASSIVE *Receiving*	ACTIVE *Manipulating*	CONSTRUCTIVE *Generating*	INTERACTIVE *Dialoguing*
LISTENING to a lecture	Listening without doing anything else but oriented toward instruction	Repeating or rehearsing; Copying solution steps; Taking verbatim notes	Reflecting out-loud; Drawing concept maps; Asking questions	Defending and arguing a position in dyads or small group
READING a text	Reading entire text passages silently/aloud without doing anything else	Underlining or highlighting; Summarizing by copy-and-delete	Self-explaining; Integrating across texts; Taking notes in one's own words	Asking and answering comprehension questions with a partner
OBSERVING a video	Watching the video without doing anything else	Manipulating the tape by pausing, playing, fast-forward, rewind	Explaining concepts in the video; Comparing and contrasting to prior knowledge or other materials	Debating with a peer about the justifications; Discussing similarities & differences

FIGURE 2.1: Examples of behaviors being classified as each of the 4 modes of engagement [8].

2.3.1 Evidence for the ICAP Framework

ICAP, although a relatively new framework, has many published papers including applications of the framework supporting its legitimacy. The first paper introducing the ICAP Framework by Chi [7] provides a brief literature review citing example studies in prior research that support the ICAP hypothesis (that learning outcomes increase as artifacts display higher levels of engagement as defined in the ICAP framework). No studies compare all four modes of engagement, but by combining many studies that all make pairwise comparisons between two of the four modes each, a case for the ICAP hypothesis is built. In total, for the 6 possible comparisons (that *interactive* is better than *constructive*, *active*, and *passive*, *constructive* is better than *active* and *passive*, and that *active* is better than *passive*) two studies were found supporting each. Also, one study for each of *interactive*, *constructive*, and *active* was found showing that different activities within the same mode of engagement produced equal results. As Chi states in the paper, the purpose of the examples are not to show exhaustively that the ICAP hypothesis is true, but to provide a starting point showing the feasibility and usefulness of ICAP.

A study by Menekse et al. [37] sought to test the ICAP hypothesis in the context of a real engineering classroom followed by a laboratory study. In the classroom study, they classified 19 activities that were already a part of the class pedagogy with minor modifications within the *interactive*, *constructive*, and *active* categories. During the first three weeks of the semester, they had students participate in activities that were a part of the three categories and take a quiz to test their knowledge at the end of class. Overall, the resulting quiz scores had significant differences between the three tested active types in favor of the ICAP hypothesis. However, when looking at the type of quiz question categorized as "integration", the *constructive* activity type produced better quiz scores than the other two

activity types. The authors theorize that this was due to the quiz question being shallower and asking for answers directly out of the instructional material.

In the same paper, Menekse et al. conducted a laboratory study with more control then the classroom study. Using textbooks on atomic bonding and physical properties they created 24 questions to be used both as a pretest and posttest. There were 15 true-false, seven multiple-choice, and two open-ended questions. There were four conditions relating to each of the ICAP modes. The *passive* condition had students students read a text passage based on the textbooks aloud without doing anything else. In the *active* condition students were instructed to highlight important sentences in the text. The *constructive* condition had students complete a graphs and figures interpretation activity, but only were able to read a shorter text than the previous two conditions. The *interactive* condition had pairs of students do the same as the constructive condition but were instructed to come to a consensus before writing their answers on a shared paper. Comparing each of the four groups pairwise showed statistically significant differences between every pair in favor of the ICAP hypothesis. The authors conclude that this is quality evidence, though they mention that this study only measured short term gains and not long term retention.

The all inclusive journal article on ICAP [8] includes the studies by Menekse and expands upon the literature review done by Chi in [7]. Throughout the literature they found the categorized learning activities such as note taking strategies and methods of forming concept maps to ICAP to then compare the performance within those studies. The paper includes many tables documenting studies that show greater learning gains for higher level ICAP categories. An example can be seen in figure 2.2. They include studies concentrating on those certain activities, as well as overall classroom studies, and conclude by saying that this empirical evidence supports the ICAP hypothesis.

	Passive	*Active*	*Constructive*	*Interactive*
Passive	No known studies			
Active	Copy map > Read map (Willerman & Mac Harg, 1991)	No known studies		
Constructive	Correcting concept map > Reading text (Chang, Sung, & Chen, 2002) Concept maps + Lecture > Lecture only for higher level material and low PK students (Schmid & Telaro, 1990) Concept maps > Study + Discussion for ESL students (Chularut & DeBacker, 2004) Concept maps > Read + Discuss (Guastello, Beasley, & Sinatra, 2000)	Building a concept map from generating > Constructing a map via selection (Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, 2005)	No known studies	
Interactive	Lecture + Collaboratively creating concept map > Lecture for science content assessment (Czerniak & Haney, 1998)	No known studies	Collaboratively building maps > individually building (Czerniak & Haney, 1998; Okebukola & Jegede, 1998)	Collaborative build concept map = Collaboratively build with 2 additional resources (van Boxtel, van Der Linden & Kanselaar, 2000)

FIGURE 2.2: Studies supporting activities from higher level ICAP categories producing more learning gains than lower level ICAP categories from [8].

These papers by Chi and Menekse et al. make a solid case for the ICAP framework being a legitimate way to categorize engagement and provide evidence that learning outcomes increase with higher engagement as hypothesized. Knowing this, the ICAP framework is a reasonable method of categorizing engagement for this thesis, but even more support comes from the large amount of research utilizing the ICAP framework in various ways.

2.3.2 ICAP in Education

In addition to these studies directly testing the feasibility of ICAP, many recent studies have been applying ICAP for different purposes throughout education research. One recent study applied the ICAP framework in the context of their nStudy learning system with the goal of proposing new learning analytics [36]. These analytics were to be for student use, so that students would be made aware of their own learning activities as opposed to an instructor learning about student activities. Additionally, a goal of these analytics was to improve the students' metacognitive skills. One of these analytics that the authors generated relies on classifying student behaviors in the system within the ICAP categories. For example, the authors describe *passive* behaviors as accessing URLs but not doing anything with the content besides reading. *Interactive* behaviors would involve using the discussion section of the system, and *active* and *constructive* behaviors would depend on whether the student's notes contain material that goes beyond the course content or not. The final analytic that the student sees displays a pie chart showing the proportion of time spent doing behaviors in each ICAP category. This is a very similar idea to this thesis, where increasing awareness about engagement during learning is one of the main end goals.

Another study was interested in the effects of *constructive* behaviors instead of *active* behaviors on understanding fractions [9]. Previous research in the domain of fractions

has shown that exposing students to multiple representations of the concept supports their conceptual knowledge. This study hypothesized that in addition to the number of representations affecting learning outcomes, students participating in constructing new fraction representations will result in more learning than only participating actively. In the *active* only condition students answered questions about fractions given to them in a web-based tutoring system. In the *constructive* and *active* condition, students split time between the web-based tutoring system and another system that allows students to explore fractions by manipulating objects, such as a jug of liquid, or sets of shapes. They found that the students who participated in *constructive* behaviors experienced a greater improvement between their pre and posttests and were also more flexible in generating fraction representations. Not only do these results back up the ICAP hypothesis, but this study is another interesting application of the ICAP framework.

2.3.3 Applying ICAP to Text Artifacts

Wang et al. studied an introductory psychology MOOC with a pre and posttest to discover whether higher quantities of participation and "higher-order" thinking behaviors are associated with higher learning gains, and then what exactly are those higher-order thinking behaviors [46]. Based in the ICAP framework, they developed a coding scheme that categorized all text artifacts as one of the four modes of engagement. Half of the data set was manually coded, and the other half used a bag of words model to automatically assign categories with around 75% accuracy. They controlled for pretest and textbook registration, but found that the *active* and *constructive* modes were most correlated with learning gains. Their explanation for the mixed results was that the posttest may have not targeted the

skills that students developed from higher-order thinking activities, and the coding scheme may have not been entirely accurate with the ICAP framework.

In their next paper, Wang et al. improved their coding scheme in the same course for further analysis [45]. They grouped students into a higher-order category of they had posted one *interactive* or *constructive* post, a paying-attention category if they had posted one active post, or an off-topic category for neither. They found that, over the entire course and controlling for the number of activities students did on the site, the higher-order category was correlated with higher course performance. For a more rigorous test, they used propensity score matching to compare pairs of students with the same activity and background but different higher-order groupings. The higher-order students had significantly, though marginally more performance. These results fall in line with the ICAP hypothesis, unlike their previous paper suggesting that the new coding rubric is more representative of the ICAP framework. Finally, they used LDA to determine what kinds of words were associated with the higher-order category. They hypothesized that higher-order texts use words connecting the material to real life situations, where other texts use more formal language from the material.

These papers by Wang et al. provide much of the basis for this thesis. They have shown that text artifacts from online lecture systems can be used with the ICAP framework and hypothesis in the same way that classroom interactions, language, and text artifacts can be used. Also, they have provided a coding rubric for text artifacts that result from discussions of lecture videos. How these fit into the methodology of this thesis is discussed in the next section where the coding rubric will be discussed further.

Chapter 3

Methodology

This chapter outlines the methodology used in this thesis. The first section describes TrACE, the system where annotations are generated and collected. Next, the process for annotation collection and coding to build data sets is described. Finally, the analyses used on the coded data sets and a description of the machine learning methods tested are discussed.

3.1 Data Collection

3.1.1 TrACE

This thesis uses TrACE, the Transformative Anchored Collaboration Environment, for data collection [16]. Instructors at multiple universities host lecture videos on TrACE and require students to view them outside of class. Often, these courses are flipped courses or taught completely online. At any point while viewing videos, students can post spatial and temporal annotations that other students will see as dots on the video. If a student clicks one of these dots, the video will pause and the annotation will be selected in the annotation bar.

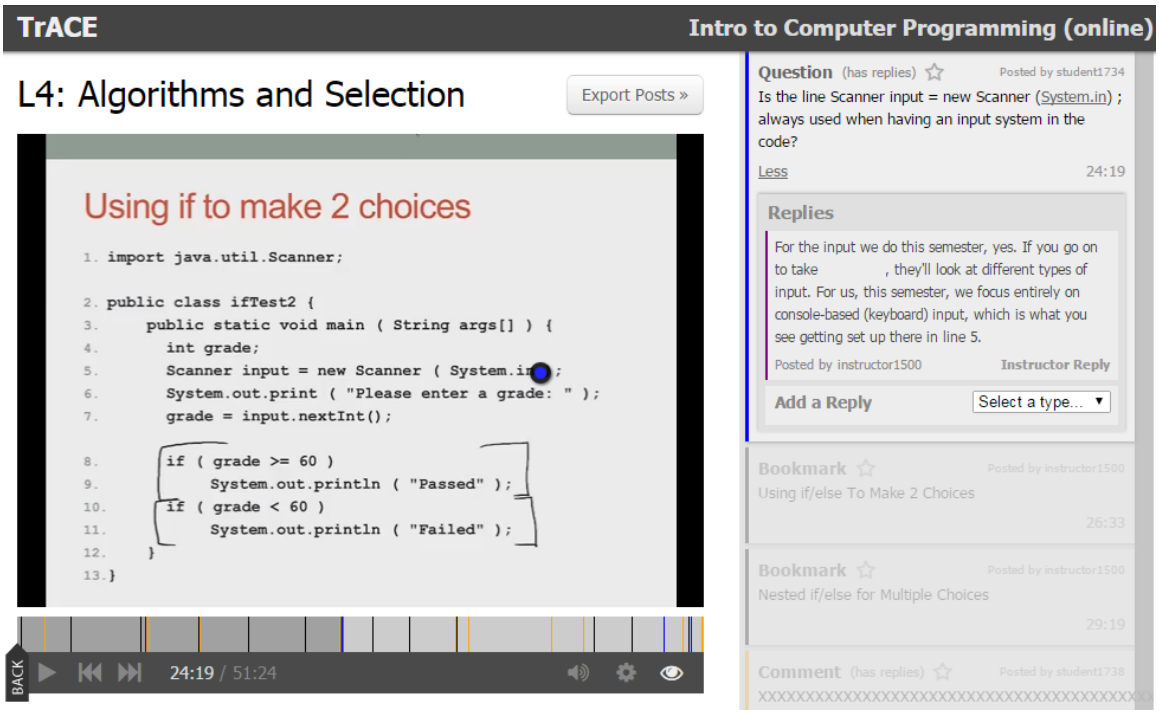

FIGURE 3.1: The TrACE video player. The annotation bar on the right hand side scrolls automatically with the video, showing annotation students have posted near where they are. The blue dot on the player is an annotation someone has posted. Students can reply to threads by selecting a reply type and typing their comment in the annotation bar.

Instructors also use these annotations to seed videos to promote discussion and reflection on video content. Annotations are stored as threads so other users can reply to allow for discussion. Figure 3.1 contains a screenshot of the TrACE video player, an annotation that someone posted, and a reply to that annotation.

Similar to other lecture video systems, click-level data about what students are doing in the system is logged, and that data is then used to produce learning analytics that give instructors insights into how students are using the lecture videos. Example actions that are logged include when students open a video, click play, pause a video, seek in a video, and leave an annotation. With this data, it is possible to recreate the stream of action students go through for behavior analyses and to calculate various metrics to be provided provided in the learning analytics. Since the lecture process occurs asynchronously in the context of

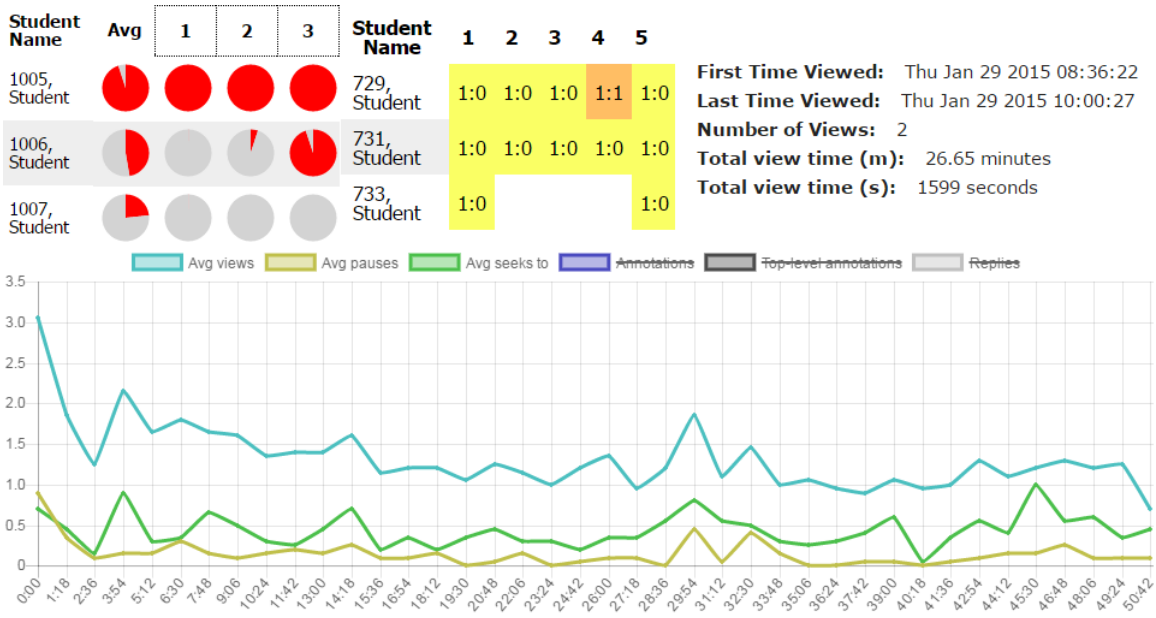

FIGURE 3.2: Examples of learning analytics within TrACE. Displayed are the percentages of each video that a student has watched (top-left), how many top-level and reply annotations a student has posted in each video (top-middle), video stats about a certain video for a student (top-right), and counts of actions plotted across the time line of a video (bottom).

TrACE (students watch at different times, and the instructor does not have any knowledge of the learning process) these learning analytics are essential to providing clues on student engagement with the lecture content to instructors. Figure 3.2 shows some of the analytics included in TrACE. As discussed in the introduction, TrACE is one of a number of research systems that use these analytics to try and recreate the affordances lost when moving from traditional lectures to online lectures.

3.1.2 Annotation Collection and Coding

To build a data sets of text artifacts, annotations from sections of the courses "Introduction to Computer Programming," "Introduction to Web Development," and "Foundations of Information Assurance" were collected. A breakdown of information about each course section can be found in table 3.1. These courses were chosen for this thesis due to being continually offered over many semesters, generating a large enough set of annotations to

TABLE 3.1: Course section information.

Course Name	Semester	Videos	Students	Total Annotations
Computer Programming	Spring 2015	26	19	282
Computer Programming	Summer 2015	26	14	404
Computer Programming	Fall 2015	26	26	728
Computer Programming	Fall 2015	26	17	370
Computer Programming	Spring 2016	26	17	507
Computer Programming	Spring 2016	26	14	279
Computer Programming	Summer 2016	26	14	323
Web Development	Spring 2015	31	11	306
Web Development	Fall 2015	31	16	315
Web Development	Spring 2016	31	17	589
Web Development	Summer 2016	31	13	274
Information Assurance	Spring 2015	10	18	116
Information Assurance	Fall 2015	10	18	133
Information Assurance	Spring 2016	10	21	154

make applying machine learning and statistical methods to them possible. Additionally, taking annotations from many sections means a wider variety of students have posted these annotations. Studying one section of students means any behaviors observed may be specific to some of those students and not generalizable to other classes or contexts.

All annotations were manually coded into the five categories summarized in table 3.2 using the same coding rubric by Wang et al. that was derived from the ICAP framework [45]. The *off-topic* category does not appear in our summary, as annotations in TrACE all refer to course content or some part of the video. Three coders were used in total, one coder working on all three courses with one coder working on one, and another working on the other two. Each course was split approximately half way between each coder working on that course. Unweighted Cohen's kappa was used to ensure intercoder reliability before coding the entire data sets. The data set for "Introduction to Computer Science" received a kappa of 0.82 on 9.0% of the data set (260 annotations), indicating strong agreement [20]. The data set for "Introduction to Web Development" received a kappa of 0.84 on 6.8% of the data set (100 annotations), indicating strong agreement. The data set for "Foundations

Table 3.2: ICAP coding categories.

Category	Description
Active (A1)	Directly mentions course content, either by asking a question about course material or repeating what is explicitly stated.
Active (A2)	Does not mention any course material, but displays attention towards the video, such as by acknowledging something said.
Constructive (C1)	Goes beyond course content by explaining or providing examples supporting ideas going beyond what is explicitly stated.
Constructive (C2)	Proposes an idea going beyond what is explicitly stated, but does not provide justification, or links to external sources.
Interactive (I)	A *C1* or *C2* behavior, but in a discussion with another student.

of Information Assurance" received a kappa of 0.80 on 12% of the data set (48 annotations), indicating strong agreement.

The rest of this thesis considers only three categories of annotations, *A1*, *A2*, and *C*, *C* including all *constructive* and *interactive* behaviors (*C1*, *C2*, and *I*). Wang et al. looked at their data this way, labeling *constructive* and *interactive* behaviors as "higher-order" behaviors, since all represent building knowledge beyond explicit course content in different ways. "Higher-order" behaviors only accounted for 12.62% of annotations, so collapsing the data in this way prevents a very small sample size for some categories precluding statistical analysis and machine learning performance evaluation. Also, this is likely the most important distinction for instructors, not the subtle differences between *C1* and *C2*. *I* behaviors, being *constructive* behaviors but in a dialogue with another student, could still be detected by checking if the text artifact was *constructive* and a reply to another student, meaning this method is still capable of distinguishing *interactive* behaviors.

3.2 Data Analysis

3.2.1 Feature Selection

To generate features for machine learning, a combination of natural language features and bag-of-words features were programmatically generated. The natural language features calculated on each annotation follow. Most features are binary (can only be 1 or 0) with the exception of length, sentences, complexity, and sentiment.

Length - How long the annotation is in words. The word count is divided by 3.

Sentences - How many sentences are in the annotation.

Link - Whether the annotation contains a URL.

Code - Whether the annotation contains common code characters not usually found in normal text ('{', '}', ';', '=', '+'). Includes HTML detection on the "Introduction to Web Development" data set. Not used on the "Foundations of Information Assurance" data set.

Question - Whether the annotation contains a question mark, indicating that the student has asked a question.

Reply to Instructor - Whether this annotation is a reply to an instructor's annotation.

Reply to Student - Whether this annotation is a reply to a student's annotation.

Complexity - The complexity of the annotation, calculated with the Automated Readability Index [44] which was designed to resemble the grade-level of the text.

Sentiment - The degree of positive or negative sentiment in the text from -1 to 1. The average is calculated from word sentiments taken from the WordNet corpus [38].

For all three data sets, bag-of-words features were selected from the most popular words found in those data sets. Bag-of-words methods simply count the number of appearances each word has in a text, and provides those counts as features. The most popular words were selected to be counted so that there were enough instances of that word for meaningful analysis. Additionally, the most popular words were selected after stop words were removed from each text, so words such as "the," "and," and "a" are not included. 300, 120, and 80 bag-of-words features were included for the "Introduction to Programming", "Introduction to Web Development", and "Foundations of Information Assurance" data sets respectively. These numbers were chosen based on a limitation of applying a MANOVA (Multivariate Analysis of Variance) to these data sets, where there needs to be less dependent variables (features) than there are samples in the smallest independent variable category. The other natural language features were selected based on the information TrACE provides and what features were seen in relevant literature (see text analyses in chapter 2). All natural language features were calculated the same way on both data sets except code.

3.2.2 Statistical Analyses and Machine Learning

One goal of this thesis is to discover what features are most useful for categorizing posts within the five engagement categories. After calculating these features for all annotations, MANOVAs are conducted on each data set with the engagement category as the independent variable and each feature as dependent variables. A MANOVA (multivariate analysis of variance) tests for a difference in means between two or more groups. In this context, given some annotations where each includes a vector of features and its category of engagement, a MANOVA tests whether the means of each feature is different when grouping annotations by level of engagement. The results of a MANOVA list, in order, which features are best at

separating annotations by engagement category. Features that were not at least marginally significant ($\alpha = 0.1$) were filtered out of the feature sets before applying machine learning. One requirement for a MANOVA is that the number of dependent variables is less than the number of data points in the smallest independent group. In this work, *constructive* annotations were the smallest engagement group for each data set, and the number of bag-of-words features for each data set were chosen based on this limitation.

The F values produced by a MANOVA describe the amount of variance between categories, but do not tell us what direction that variance goes. For example, a high F value for the *length* feature would indicate that *length* has a high level of variance between engagement category groups, and is a good feature for machine learning, but does not tell us whether *C* annotations are generally longer than *A1* categories or vice versa. To perform this sort of analysis, separate ANOVAs are run with each feature as the dependent variable and ICAP category as the independent variable for features that were significant ($\alpha = 0.05$) in the MANOVA. A post-hoc Tukey's Honest Significant Test (TukeyHSD) is then conducted to calculate the variance between each pair of engagement categories (*A1* with *A2*, *C1* with *A1*, and *C1* with *A2*), enabling an analysis of how these features affect engagement category. Another TukeyHSD test looks to see if the means are significantly different between posts in one engagement category between courses, showing if this machine learning model can be easily generalized to many courses.

Finally, performance of a machine learning algorithm was tested on the programming and web development data sets using all of the marginally significant features from the MANOVAs. The information assurance data set was not included in machine learning testing due to the much smaller sample of annotations, making machine learning impractical. The machine learning algorithm used is a multilayer perceptron, a type of neural network

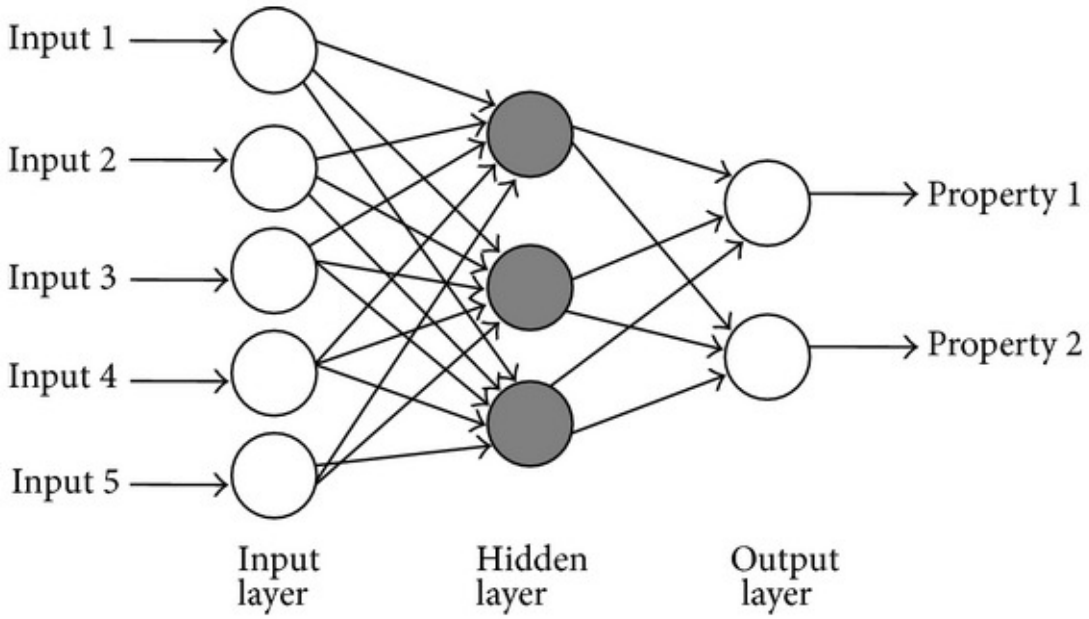

FIGURE 3.3: The connections of a multilayer perceptron, where there can be many hidden layers between the inputs and outputs [14].

that in this case is trained to map features to engagement category. Between the input layer and output layer are one or more hidden layers of nodes or neurons. Figure 3.3 contains a diagram of a multilayer perceptron with one hidden layer. The multilayer perceptron was chosen based on its presence in related work and the success of neural networks in language learning problems.

3.3 Summary

This study collects all data from TrACE, a website where instructors host lecture videos for students who watch and annotate them. Annotations from sections of "Introduction to Computer Programming," "Introduction to Web Development," and "Foundations of Information Assurance" were collected from TrACE. Using a coding rubric based on the ICAP framework, annotations were manually coded based on the level of engagement with

course material that was displayed. With each data set, a MANOVA was run to filter the feature sets, pairwise comparisons were run to observe the differences between engagement categories, and the filtered feature sets were used to test the feasibility of a multilayer perceptron automatically categorizing new annotations. In the last step, sketches of what an implementation of a learning analytic using this automatic categorization of new annotations were made based on previous research on TrACE instructor behaviors. The next section begins the results of these methods.

Chapter 4

Results and Discussion

In this chapter, the results from all of the methods in the previous chapter are provided with notable findings highlighted. First, the MANOVA results for each dataset are shown to help answer the first research question, what features are most important for classification. Next, pairwise comparisons of each feature from post-hoc TukeyHSD tests are given helping answer the second research question of whether a machine learning model can be generalized to different classes. Then, the results of more TukeyHSD tests are shown to test the variance in features between courses, further answering the second research question. Lastly, the results of testing a multilayer perceptron on automatically classifying annotations with the filtered list of features is shown to answer weather applying machine learning in this context is feasible. The results from this chapter also help inform the design of learning analytics that would use such a system in the next chapter.

4.1 MANOVA

The first statistical method used on all data sets is the MANOVA, which compares the variance between means of features when grouped by engagement category. The purpose

TABLE 4.1: MANOVA F values for natural language features where $p<0.05$. Features are ordered by programming F value. Rank indicates that feature's place in all features of that data set ordered by F value.

Feature	Programming F	Rank	Web Dev. F	Rank	IA F	Rank
Length	455.26	1	168.28	1	76.79	1
Complexity	323.20	2	138.37	2	68.34	2
Question	279.86	3	83.03	4	6.67	18
Reply to Inst.	132.36	4	114.31	3	5.32	23
Code	115.27	5	15.39	19		
Link	78.245	7	53.43	5	29.28	4
Reply to Stu.	70.81	8	17.70	15	15.89	10
Sentences	44.929	15	35.01	6	32.571	3
Sentiment	34.853	22	16.62	17		

of this test is two-fold; It shows what features had significant differences between means, indicating that that feature is useful for machine learning, and also shows which features had the highest variance in means through the F value. The results from the natural language features are given first, and the bag-of-words features second.

4.1.1 Natural Language Features

Table 4.1 contains the F values from each MANOVA for the natural language features, ordered by the F values from the programming data set. Missing numbers indicate that the means of that feature was not significantly different between engagement category. This occurs with the code and sentiment features on the information assurance data set, the code feature explained by the lack of coding in that course.

The length feature had the highest F value for all three data sets, followed by the complexity feature, indicating that these features have the most varied means between engagement categories. Intuitively, this result suggests that students who put more effort into their annotations, resulting in longer and more complex language, results in higher engagement. Similar results correlating length and complexity with engagement have been

found in prior work [2, 32]. Furthermore, almost all works discussed in the related works section that looked at performance as compared to activity behaviors in learning systems found correlations between effort and performance. Pairwise analyses in the next section will discover whether it is indeed higher engaged annotations that have higher length and complexity.

Largely, the programming and web development data sets display similar patterns in terms of their feature orders when ranked by F values. The main differences come from the code and reply to student features which had very different rank placements. The web development course had a much smaller emphasis on code, and many lectures discussed web technologies. On the other hand, the introduction to programming course focused almost entirely on coding, and videos largely consisted of programming examples. The change in code rank then makes sense, as students writing code in their annotations suggests that they are asking more detailed questions than those who are not, indicative of higher engagement.

In the information assurance data set, question, reply to instructor, and reply to student scored much lower than in the other two data sets. One hypothesis as to the discrepancy between the reply features is that the instructor for the information assurance course included question annotations that students were supposed to answer as they watched the video. The other two courses did not have these questions, only general comments by the instructor. Students then in the information assurance course were likely to respond to these instructor posts with simple answers to these questions without detailed explanation resulting in lower engagement categories.

Despite some differences observed between the data sets, almost all of the natural language features had significantly varied means between engagement categories, with the exception of code (expected) and sentiment on the information assurance data set. All

of these features will then be included as features for the machine learning performance evaluation and the programming and web development data sets. Having chosen these features based on natural language analyses in previous research looking at student success, these results provide evidence that these engagement categories may also be correlated with student success.

4.1.2 Bag of Words Features

Table 4.2 contains the F values resulting from the MANOVAs for the top 10 bag-of-words features of each course ranked by F value. Looking at these words, many of them refer to specific topics that are discussed in their courses such as "array," "variable," and "int" for the programming course, "html," "css," and "perl" for the web development course, and "access," "passwords," and "com" for the information assurance course. This relationship between course "keywords" and engagement category is expected, as students who are more engaged with the course material use these words to discuss course topics. Students posting off topic or simple questions will likely not refer much to these keywords, resulting in a lower engagement category. To see if this is the case however, pairwise comparisons between groups in the next section are needed to see if it is actually the more engaged annotations that are using these words more often.

Interestingly, the word "would" appears in the programming and web development data sets, as well as the word "understand" and "use" or "using" in the web development and information assurance data sets. These words are the only words in the top 10 that have no direct relationship with the course material. What exactly the relationship between these words and engagement is will be teased out with the next statistical analysis.

TABLE 4.2: MANOVA F values for bag-of-words features where p<0.05. Shown are the 10 words with the highest F values for each course.

Progamming Word	F	Web Dev. Word	F	IA Word	F
would	79.40	html	29.63	video	28.79
array	60.06	css	26.17	access	21.90
variable	58.83	perl	23.74	passwords	18.88
int	56.82	way	22.60	com	17.76
java	55.68	web	19.05	password	15.95
arrays	55.30	chrome	18.33	attacker	10.17
method	53.06	color	18.01	https	9.26
value	43.03	use	17.77	using	9.13
number	42.96	would	17.42	understand	9.12
material	40.84	understand	15.85	hash	8.08

4.2 Pairwise Comparisons

Beginning pairwise comparisons between engagement categories, the following results come from running an ANOVA with a feature as the dependent variable and engagement category as the independent variable with a post-hoc TukeyHSD test. The TukeyHSD test runs pairwise comparisons to test the variance between engagement categories, providing insight into whether features are indicative of engagement or disengagement. By observing the direction of these differences for the same feature in different courses, the last research question of whether this method can be generalized to all courses can be better answered. The results from the natural language features are given first, and the bag-of-words features second.

4.2.1 Natural Language Features

Table 4.3 contains the results from running post-hoc TukeyHSD tests on the natural language features. For each natural language feature, all differences between engagement groups that were significant have the same polarity between courses. For example, the differences show that a longer length means the annotation is more likely to be of higher

engagement for all courses, as the differences are all positive. On the other hand, an annotation being a reply to an instructor's annotation means that it is more likely to be of lower engagement for all courses, as the differences are all negative. This result is evidence that the method used in this thesis to automatically categorize annotations by engagement can be generalized to more courses. If the polarity of some features were reversed between features in different courses, that would mean that engagement manifests itself in different ways through text artifacts in those courses. Then, it would be necessary for a machine learning algorithm to be trained and used on only that course. This is not the case here, and suggests that engagement manifests itself in similar ways through text artifacts in TrACE, at least in the courses studied in this thesis.

The only natural language features that have inverse relationships with engagement are reply to instructor, sentiment, and reply to student between *A1* and *A2* posts, where all other natural language features are indicative of higher engagement. To explain the inverse relationship with instructor replies, often instructors post annotations containing simple questions to try and make students reflect on the video material. Students responding to these annotations often answer the question, without constructing any new knowledge, resulting in lower engagement levels for those student annotations. On the other hand, student replies result in higher engagement, except between *A1* and *A2* posts. This means that *C* has the highest mean for the reply to student feature, followed by *A2* and then *A1*. One hypothesis for this result is that if a student is replying to another student, they are doing so either because they are building upon a question they had (resulting in a *C* post), or for meta discussion about the course or video resulting in *A2* posts (meta discussion examples from the programming data set include "In the long run, classes like these are great...," "Kind of. There is just so much information in this course!" and "I didn't even

TABLE 4.3: Pairwise TukeyHSD results for natural language features. "Diff" columns show, for a feature and course data set, the variance between means of that feature between the specified engagement groups. Missing values indicate that the variance in means between those two engagement groups was not significant ($\alpha = 0.05$).

Course	Feature	A1-A2 Diff	C-A1 Diff	C-A2 Diff	MANOVA Rank
Programming	Length	1.157	2.396	3.553	1
Web Dev.	Length	1.703	1.436	3.139	1
Info. Assurance	Length		5.080	6.000	1
Programming	Complexity	3.416	6.804	10.221	2
Web Dev.	Complexity	5.055	3.641	8.697	2
Info. Assurance	Complexity		14.006	17.330	2
Programming	Question	0.245	0.225	0.470	3
Web Dev.	Question	0.182	0.159	0.340	4
Info. Assurance	Question	0.216			18
Programming	Reply to Ins.	-0.158	-0.293	-0.451	4
Web Dev.	Reply to Ins.	-0.180	-0.443	-0.623	3
Info. Assurance	Reply to Ins.	-0.166		-0.255	23
Programming	Code	0.047	0.153	0.200	5
Web Dev.	Code	0.051		0.088	19
Info. Assurance	Code				
Programming	Link		0.059	0.059	7
Web Dev.	Link		0.088	0.088	5
Info. Assurance	Link		0.167	0.167	4
Programming	Reply to Stu.	-0.031	0.177	0.147	8
Web Dev.	Reply to Stu.	-0.055	0.133	0.079	15
Info. Assurance	Reply to Stu.		0.190	0.157	10
Programming	Sentences	0.123	0.281	0.404	15
Web Dev.	Sentences	0.319		0.416	6
Info. Assurance	Sentences		1.026	0.912	3
Programming	Sentiment	-0.058	-0.052	-0.110	22
Web Dev.	Sentiment	-0.034	-0.101	-0.136	17
Info. Assurance	Sentiment				

notice that 'till you pointed it out."). Students may be unlikely to respond to another student with an *A1* post that does not build upon what the first student said.

Another interesting result is that the sentiment feature had an inverse relationship with engagement for all engagement categories. Although the variance in means is small (0.034 to 0.110, where the sentiment feature is bounded by -1 and 1), this result shows that higher engaged annotations have a slightly more negative sentiment. This may be explained by the affective states of learning discussed in the introduction. If a student is in the *satisfied* state, being confident that they have a mastery of the video material, they will not post any questions or attempt to resolve any confusion. Likewise, students experiencing the *hopefulness* state are unlearning their misconceptions and confident that they are on their way to learning. The students experiencing negative affective states (*confusion* and *frustration*) may be the ones posting more engaged annotations in an attempt to ask questions, construct knowledge, and unlearn misconceptions by engaging with peers and the instructor. In related work studying MOOC attrition with sentiment analysis, a Python course found that very positive or very negative posts were correlated with dropping out of the MOOC [48]. That result always plays into the narrative of affective states, where students experiencing higher amounts of frustration post very negative posts, and unable to overcome the frustration drop out of the MOOC. In the context of TrACE, this could mean that students somewhere in the middle are those posting higher engaged posts, with those experiencing a more negative state posting questions to resolve that confusion. Although interesting, future work with sentiment in this context is needed to further study this phenomenon.

4.2.2 Bag of Words Features

Table 4.4 contains the results of running post-hoc TukeyHSD tests on the 10 highest ranked words from the MANOVA results for each course. One interesting result is that most courses "keywords" correlate with higher engagement, with the exception of "arrays" between *C1* and *A1* in the programming data set.

The word "understand," which appeared in the web development and information assurance data sets, has an inverse relationship with engagement, as well as "video" in the information assurance data set and "material" in the programming data set. Although discovering the reason for these relationships requires further research, we can conclude that these words are used more often in less engaged posts. One explanation for the word "understand" could be that students often use the word when they are simply reporting whether they understood the material, or asking a simple question while saying that they do not understand something as in an *A1* post. The words "video" and "material" may have an inverse relationship since they are commonly used for meta discussion, such as through suggesting improvements to the video, or reporting that they did or did not understand the video material. Finally, the only non-keyword feature to have a positive relationship with engagement was "would," which even had the highest F value of bag-of-words features in the programming data set. It could be that the word "would" is often used in hypothetical questions, often resulting in higher levels of engagement an students are trying to construct knowledge. With all of these words, more qualitative research methods are needed to discover why these relationships with engagement exist.

To summarize this section, the results from these pairwise comparisons is evidence that the methods used to train a machine learning algorithm to automatically classify annotations by engagement may be generalizable to different courses. All of the natural

TABLE 4.4: Pairwise TukeyHSD results for the top 10 bag-of-words features. "Diff" columns show, for a word and course data set, the variance between means of word occurrences between the specified engagement groups. Missing values indicate that the variance in means between those two engagement groups was not significant ($\alpha = 0.05$)

Course	Word	A1-A2 Diff	C-A1 Diff	C-A2 Diff	MANOVA Rank
Programming	"would"	0.072	0.195	0.267	6
Programming	"array"	0.048	0.120	0.168	9
Programming	"variable"	0.016	0.088	0.104	10
Programming	"int"		0.098	0.112	11
Programming	"java"	0.023	0.108	0.131	12
Programming	"arrays"	0.118	-0.051	0.067	13
Programming	"method"	0.057	0.118	0.175	14
Programming	"value"	0.028	0.070	0.098	16
Programming	"number"	0.028	0.074	0.102	17
Programming	"material"	-0.077	-0.045	-0.122	18
Web Dev.	"html"	0.111		0.111	7
Web Dev.	"css"	0.095		0.055	8
Web Dev.	"perl"	0.089		0.080	9
Web Dev.	"way"	0.033	0.077	0.120	10
Web Dev.	"web"	0.053	0.060	0.113	11
Web Dev.	"chrome"		0.075	0.088	12
Web Dev.	"color"		0.068	0.083	13
Web Dev.	"use"	0.099		0.110	14
Web Dev.	"would"	0.081		0.144	16
Web Dev.	"understand"	-0.074		-0.140	18
Info. Assurance	"video"	-0.455		-0.520	5
Info. Assurance	"access"		0.317	0.382	6
Info. Assurance	"passwords"		0.394	0.431	7
Info. Assurance	"com"		0.133	0.137	8
Info. Assurance	"password"		0.465	0.598	9
Info. Assurance	"attacker"		0.164	0.225	11
Info. Assurance	"https"		0.115	0.127	12
Info. Assurance	"using"		0.143	0.196	13
Info. Assurance	"understand"	-0.116		-0.157	14
Info. Assurance	"hash"	0.078	0.197	0.275	15

language features had the same relationship with engagement in each course. Although this thesis uses bag-of-words features, which have been shown to be course specific, the most important features have been the natural language features. Furthermore, a clear pattern from the pairwise analysis of bag-of-words features shows that course "keywords" are the important words, and that they have positive relationships with engagement. Given a list of course keywords, a system could count only them for bag-of-words features or count a single keyword feature for classification in place of the bag-of-words method used in this work. Likely, having seen the much greater F values for the natural language features and keywords making up the vast majority of the top 10 words, performance would not be hindered using this keyword method. The next section includes an additional analysis of features to more concretely answer the second research question.

4.3 Between Course Comparisons

In the previous section, the TukeyHSD results comparing the variance in means of features between engagement category showed that the natural language features, as well as course "keywords" in the bag-of-words features, all had the same relationships with engagement. This is evidence that the method for classifying student posts by engagement is able to be generalized to many courses, but for a machine learning model to be trained on one course and applied to a different course would require little variance in the means of features. Table 4.5 shows the significant results from running the same TukeyHSD test as the previous section, but comparing the means of features for some engagement category with those from a different course. For example, the first row in the table shows the differences in the length feature for *A2* posts between the programming, web development, and information assurance data sets. There was not a significant variance in the length feature between

TABLE 4.5: TukeyHSD results comparing feature means for engagement categories between courses. "Diff" columns show the variance in means of that feature between courses. Missing values indicate that the variance in means between those two courses was not significant ($\alpha = 0.05$). The code feature was not used in the information assurance data set resulting in missing values for that feature.

Category	Feature	Prog.-Web. Diff	Prog.-Info. Diff	Web.-Info. Diff
A2	Length		1.844	1.786
A1	Length	0.603	1.603	1.000
C	Length		4.291	4.648
A2	Complexity		5.294	5.423
A1	Complexity	1.505	5.202	3.697
C	Complexity		12.404	14.057
A2	Question	0.281		-0.264
A1	Question	-0.073		
C	Question	-0.139	-0.335	-0.196
A2	Reply to Instructor	0.188	0.311	
A1	Reply to Instructor	0.166	0.303	
C	Reply to Instructor		0.507	0.491
A2	Code			
A1	Code			
C	Code	-0.108		
A2	Link			
A1	Link			
C	Link		0.108	
A2	Reply to Student	0.030		
A1	Reply to Student			
C	Reply to Student			
A2	Sentences	0.083	0.346	0.263
A1	Sentences	0.279		-0.170
C	Sentences		0.854	0.759
A2	Sentiment	0.041		-0.119
A1	Sentiment	0.065	-0.064	-0.129
C	Sentiment			

the programming and web development data sets for *A2* posts. However, programming *A2* posts and web development *A2* posts were significantly longer than information assurance *A2* posts.

Many features had significant variances in means between courses for some engagement

categories, indicating that more work is likely needed to generalize the machine learning method in this thesis to many courses. *C* posts in the programming and web development data sets were on average around 13 words longer than *C* posts in the information assurance data set. Complexity, which when calculated with the Automated Readability Index is supposed to resemble the grade-level of a text [44], was much lower in *C* posts for the information assurance data set than the other two data sets. These results show that students in the information assurance course were making *C* posts that were much shorter and less complex than those in the other two courses. Another major difference can be seen in the reply to instructor feature. One of the largest differences comes from *C* posts in the programming and web development data sets being significantly more likely to be replies to instructors than *C* posts in the information assurance data sets. There were also significant differences in the reply to instructor feature for *A2* and *A1* posts between programming and the other two data sets. Mentioned here have been some notable findings, though there are many more significant differences to be seen in table 4.5.

The larger implication of these significant differences is that a machine learning model trained on one course would likely have poor accuracy when applied to a different course, especially with the information assurance course having much lower complexity for all types of posts. Discovering the reason behind these differences is outside the scope of this thesis, but future work would need to consider these findings when attempting to develop a similar machine learning model for many courses. Knowing this, the next section looks to test the viability of this model by training and testing on data within courses.

4.4 Machine Learning

Table 4.6 contains metrics from the performance evaluation of a multilayer perceptron categorizing annotations based on engagement category. Again, the information assurance course was not included here due to the much smaller sample size making machine learning applications infeasible (403 annotations versus 1484 in the web development data set and 2893 in the programming data set). The metrics were calculated by testing and training on 10 randomly stratified samples where 15% of the data was chosen as the test set and the rest for training. Samples were stratified by grouping annotations by what video they were posted in, and then selecting 15% of annotations from each video for the test set. For example, each test set contained 15% of the annotations from video 1, 15% of the annotations from video 2, 15% of the annotations from video 3 etc. This stratification was done to ensure that the train and test sets were representative of discourse throughout the whole course, as the content being discussed may influence behavior. Bag-of-words features specifically will change depending on the video an annotation is in because of the new material and course keywords being discussed.

The performance evaluation of a multilayer perceptron automatically categorizing annotations by engagement category resulted in good performance overall, achieving 80.03% accuracy on the programming data set and 74.06% on the web development data set. Classification of *A1* and *A2* posts performed well with precision and recall metrics all over 70%. However, precision and recall for C posts leaves room for improvement. Specifically, recall for C posts was very poor on the web development data set with an average of 25.25%. One cause of this may be the much smaller number of C posts in the web development data set, only 146 where the programming data set had 349. Splitting 146 annotations into test and train sets means that any machine learning algorithm might not have enough data for

TABLE 4.6: Classifier metrics, averaged from 10 randomly stratified samples.

Metric	Programming Dataset	Web Development Dataset
Overall Accuracy	**80.03% (σ = 1.95%)**	**74.06% (σ = 1.68%)**
A1 Precision	77.81% (σ = 3.77%)	71.39% (σ = 4.82%)
A1 Recall	78.18% (σ = 4.47%)	79.20% (σ = 3.50%)
A2 Precision	83.44% (σ = 3.10%)	78.29% (σ = 3.51%)
A2 Recall	90.04% (σ = 2.02%)	79.87% (σ = 4.64%)
C Precision	76.54% (σ = 11.71%)	64.35% (σ = 17.23%)
C Recall	52.45% (σ = 7.60%)	25.25% (σ = 7.00%)

a full understanding of C behavior to categorize new C posts. The significantly improved metrics on the programming data set suggests that this may be the case.

To answer the last research question, whether it is feasible to use a machine learning algorithm to categorize posts by engagement category, the data suggests that it is indeed possible. With the exception of poor performance with C annotations on the web development data set, all performance metrics suggest that the methods used to run this algorithm provide enough accuracy to be used in a real-life environment with learning analytics. Larger systems with data from a greater number of students would be able to overcome the small proportions of C posts that seem to be affecting performance in this thesis. Furthermore, having found that generalizing the methods used in this thesis to all courses is likely possible, future research could combine the data sets for each course to remove this issue of sample size.

4.5 Summary

In this chapter, the results of three different analyses were presented to answer the three research questions posed in the introduction. First, MANOVAs were run on each data set to determine which features were most important for classification. It was found that all of the natural language features selected were significant indicators of engagement, though

sentiment was by far the least important. Looking at the bag-of-words features, course keywords, such as "variable," "java," and "arrays" in the programming data set, were found to be the most important classification. Only a few words, "would," "understand," "use," and "video," made it to the top 10 of some data sets and were not a word related to course content. Next, pairwise analyses were conducted to answer to help answer the second research question, whether the machine learning methods in this thesis can be generalized to more courses. Based on all natural language features and course keyword features having the same relationship with engagement (with the exception of "arrays" on the programming data set), generalizing to multiple courses seems possible. However, the next pairwise analysis showing that means are often significantly different between posts in different courses shows that the model in this thesis would need to be modified before generalizing it is possible. Lastly, all marginally significant features from the MANOVA results were used to train and a test a multilayer perceptron for automatic categorization of engagement. Performance metrics indicate that this method is feasible, and better performance can theoretically be achieved using a larger data set. In the next chapter, these results and previous research instructors are combined to generate mockups of what a learning analytic built on this machine learning method would look like.

Chapter 5

Implications for Design

This chapter utilizes the results from the previous chapter, in addition to prior research on instructor's use of TrACE and its analytics, to create mockups of an analytic utilizing the proposed machine learning system that categorizes engagement. In previous work studying the role of formative assessment in the classrooms of TrACE instructors, journals written by instructors and interview transcripts were used in a thematic analysis to identify themes in instructors experiences of TrACE [18]. Table 5.1 shows the themes discovered via superordinate and subordinate categories. Following are a subset of those themes that help inform the design of learning analytics that would use the machine learning method described in the previous chapter to provide student engagement data to instructors. Each category is paraphrased here. For full descriptions see [18].

- **Knowledge of Students**: This first superordinate category was chosen due to instructor often speaking about their awareness of student activity and how they conceive student knowledge and understanding. The two subordinate categories described next show instructor desires that influenced the topic of this thesis, specifically working towards a student engagement learning analytic.

TABLE 5.1: The superordinate and subordinate themes from research into how instructors apply formative assessment within TrACE.

Superordinate Categories	**Subordinate Categories**
Knowledge of Students	Engagement
	Viewing Patterns
	Understanding
	Experiencing Difficulty
Quality of Materials	–
Educator Action	Individual Student Intervention
	Group Action or Instructional Change
	Next Course Iteration
General Limitations	System Shortcomings
	Educator Struggles

  - **Engagement**: This subordinate category describes instructor's use of analytics to learn how students are engaging with the video content. On the macro level, instructors used both analytics and annotation summaries in TrACE to discover suspicious activity and whether students are meeting course expectations. At the micro level, instructors used finer grained analytics that describe their actions in the system to see if students were focusing on the important content and not cherry picking parts of the video.
  - **Experiencing Difficulty**: Instructors, in addition to wanting to know how students use TrACE, sought to identify confused or struggling students. One instructor used analytics showing where students made actions in the video to get an idea of what concepts students were struggling with. A common sentiment was that the analytics were not enough to judge whether students were struggling.

- **Educator Action**: After instructors acquired knowledge of students through themes in the first superordinate category, the next step for instructors was to put that

information into action.

- **Individual Student Intervention**: One instructor showed an analytic of a students behavior to that student to open a discussion on why the instructor is concerned about the student's progress. Other instructors also used analytics to identify students needing intervention to try and get them engaged in the course.
- **Group Action or Instructional Change**: In addition to targeting individual students, instructors also addressed the course as a whole when engagement seemed low, and altered the course in some way such as by changing the requirements for using TrACE. One instructor would give quizzes to students who did not watch the entire video for that day, and another discussed common questions they saw in TrACE.

5.1 Analytics on Classes

The two analytic mockups in this section focus on analytics that provide information about the class as a whole. Instructors will not be able to discern specific information about students from these, but may be able to identify general class trends for group action or instructional change.

Figure 5.1 displays a mockup of an analytic that aims to show how post counts in each engagement category change over time. In the analytic is a graph that displays the total number of annotations in each engagement category for each video in the course. This analytic mainly addresses the needs found in the subordinate category "Group Action or Instructional Change." Instructors, often looking at analytics to observe the overall engagement levels of the class, may use this information to make instructional change that

then promotes engagement. Although there are currently analytics in TrACE that describe engagement through log data (such as the number of annotations posted in a video by each student) over all videos in a course, instructors need to look at many analytics to build an overall picture of engagement. Using levels of engagement instead of log data also provides a more accurate depiction of how engaged the student is with the course material since log data may not be representative of student understanding.

One possible use case of this graph is an instructor looking to see how engagement levels change throughout the course. If half way through the course the levels of engagement begin to dip, it may mean that the instructor's expectations are not in line with what students believe the expectations are. The instructor may then intervene or change course policy to remedy the problem. This analytic might also be used to identify what videos students most struggle with. Lower levels of engagement may suggest that students are not connecting with the material, and negative affective states are resulting in disengagement. If a video consistently has lower engagement than the others over multiple iterations of a course, it may mean that students need more help with that material, or that the video production is of less quality than the others resulting in disengagement [23].

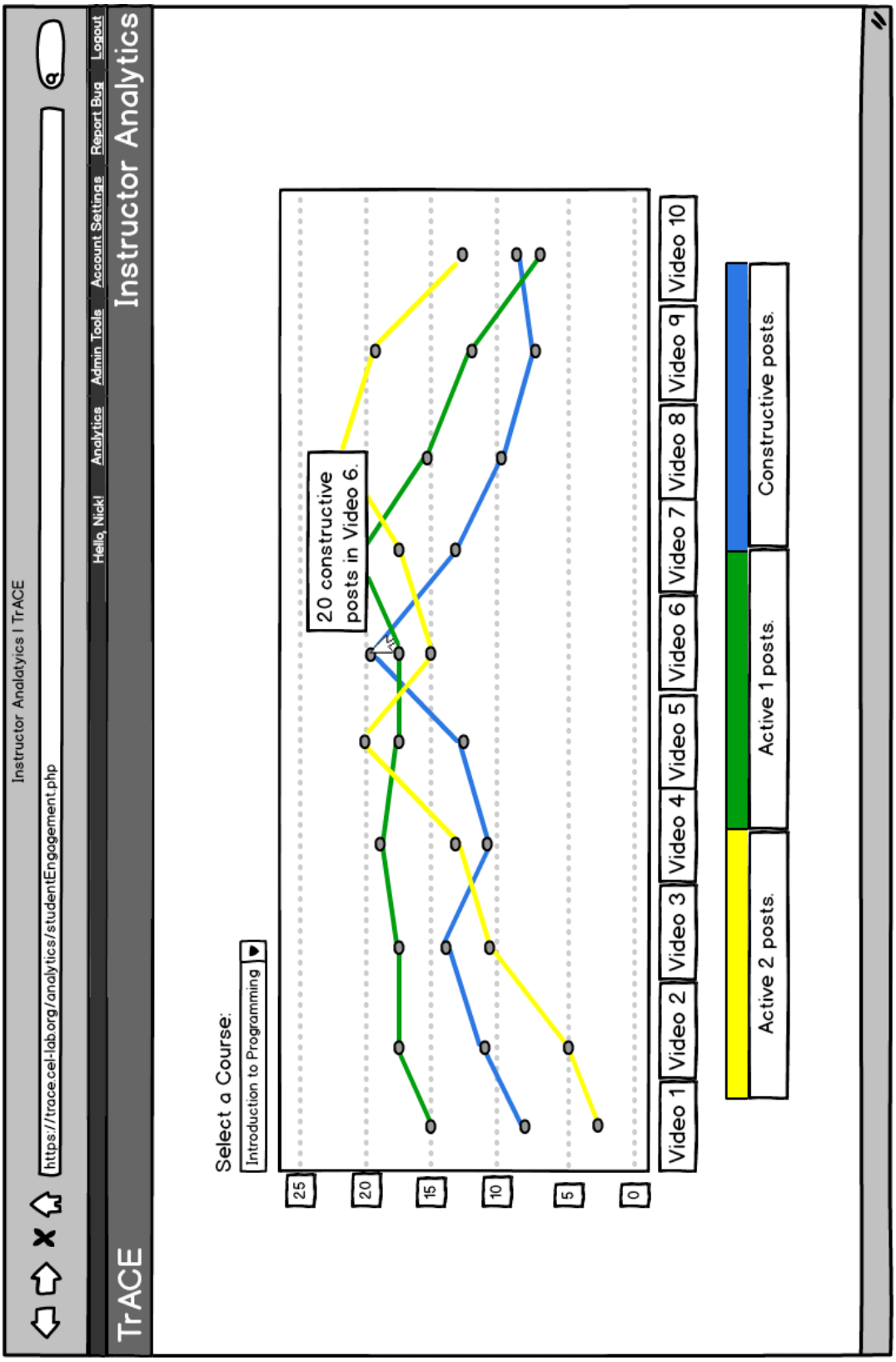

FIGURE 5.1: A class engagement graph by video.

Figure 5.2 shows an analytic with a different type of graph, one that displays the number of annotations in each engagement category at different times in the video. As annotations in TrACE are anchored in time and pixel coordinates on the video, looking at the annotations in this way shows which annotations correspond to what course material based on what time the annotation is posted at. This most helps the category "Engagement," where instructors wanted to discover suspicious activity and whether students were cherry picking content and not paying attention to all of the material. This graph enables instructors to do so by allowing them to see trends in where engagement is and is not occurring within the video. Spots of no annotations or low engagement annotations may indicate that students need help with that section of course content.

Another possible use case is using the changes in engagement levels to help discover what videos or portions of the course students struggle with the most, whether this be due to the difficulty of the material or video production. Previous research looking at log data in a MOOC found that the type of video segment influenced behavior, and that looking at engagement data can help inform instructors about where videos can be improved [23]. They also developed a list of video production practices based on instructor interviews that promote student engagement [22]. Knowing this, instructors using this graph may be able to identify where videos can be improved for future iterations of the course.

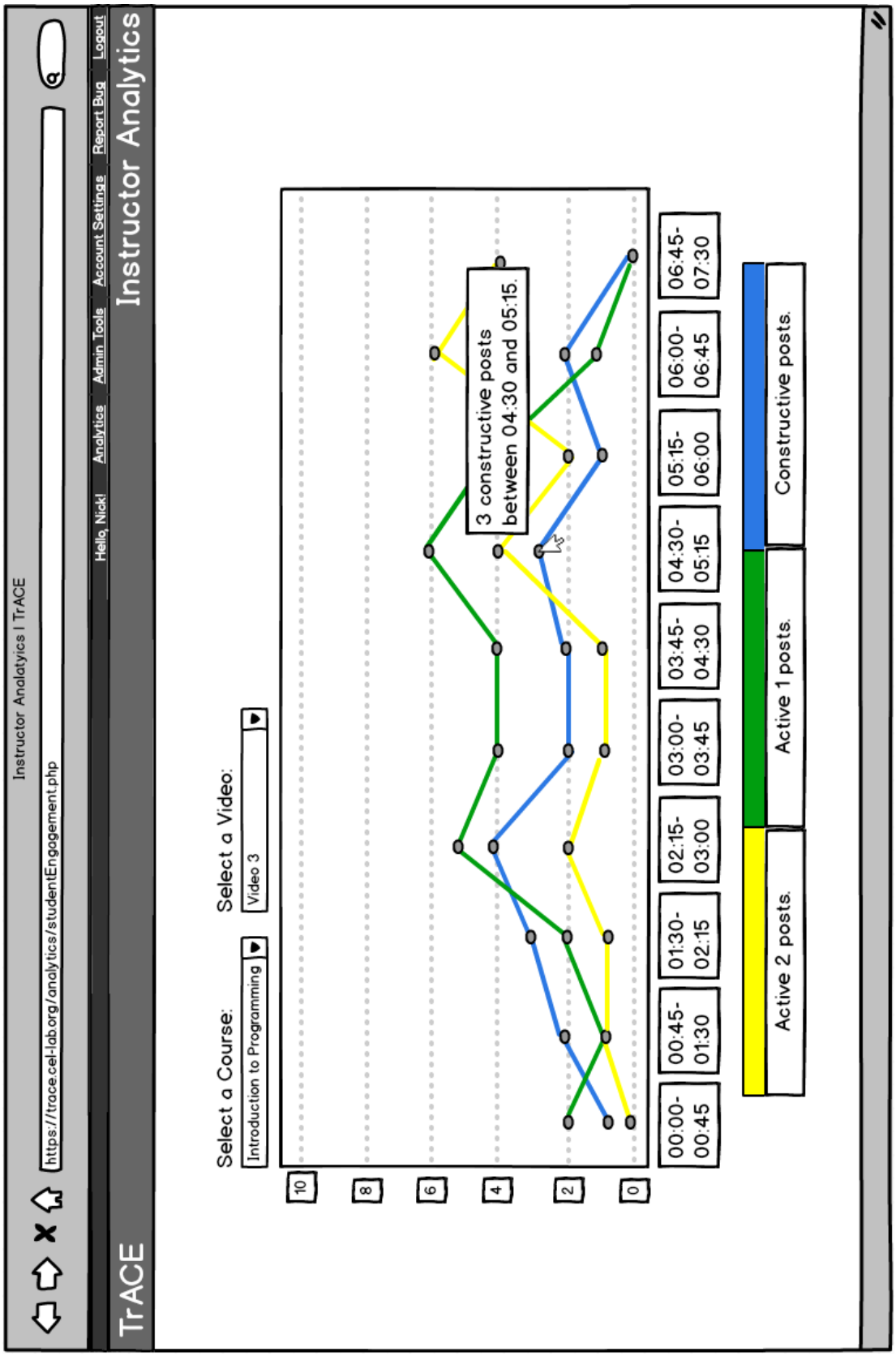

FIGURE 5.2: A class engagement graph by time in a video.

5.2 Analytics on Students

The two analytics in this section focus on providing information to instructors about individual students. Where the previous two analytics allowed instructors to observe general class trends, these will enable instructors to see trends in each student's behavior. These analytics more directly address issues with online learning discussed in the introduction: that instructors are unaware of student engagement during the learning process, are unable to intervene students, and are not able to scaffold learning for students to move them through the affective states of learning which may disproportionately affect weaker students.

The analytic in figure 5.3 seeks to address needs from the categories "Experiencing Difficulty" and "Individual Student Intervention." Displayed is a chart with videos on the x-axis and students on the y-axis, which each box representing the number of annotations in an engagement category that student has posted. The current engagement category can be changed by the drop-down box at the top, which in the figure is set to *constructive*. As it is in the figure, the number of *constructive* annotations each student has posted in each video is being displayed.

A possible use case for this analytic is that by observing which students consistently post no constructive annotations in videos, instructors may be able to identify students who are struggling with keeping up in the course. Also, changing the type of annotations displayed to *active 2* may show students who are posting annotations as the instructor expects, but are consistently not using those annotations to discuss course content. The instructor then may use other analytics to see what the student is using the annotations for, and then may decide if individual student intervention would be beneficial. This analytic overall aids instructors in discovering how engaged each student is with course content.

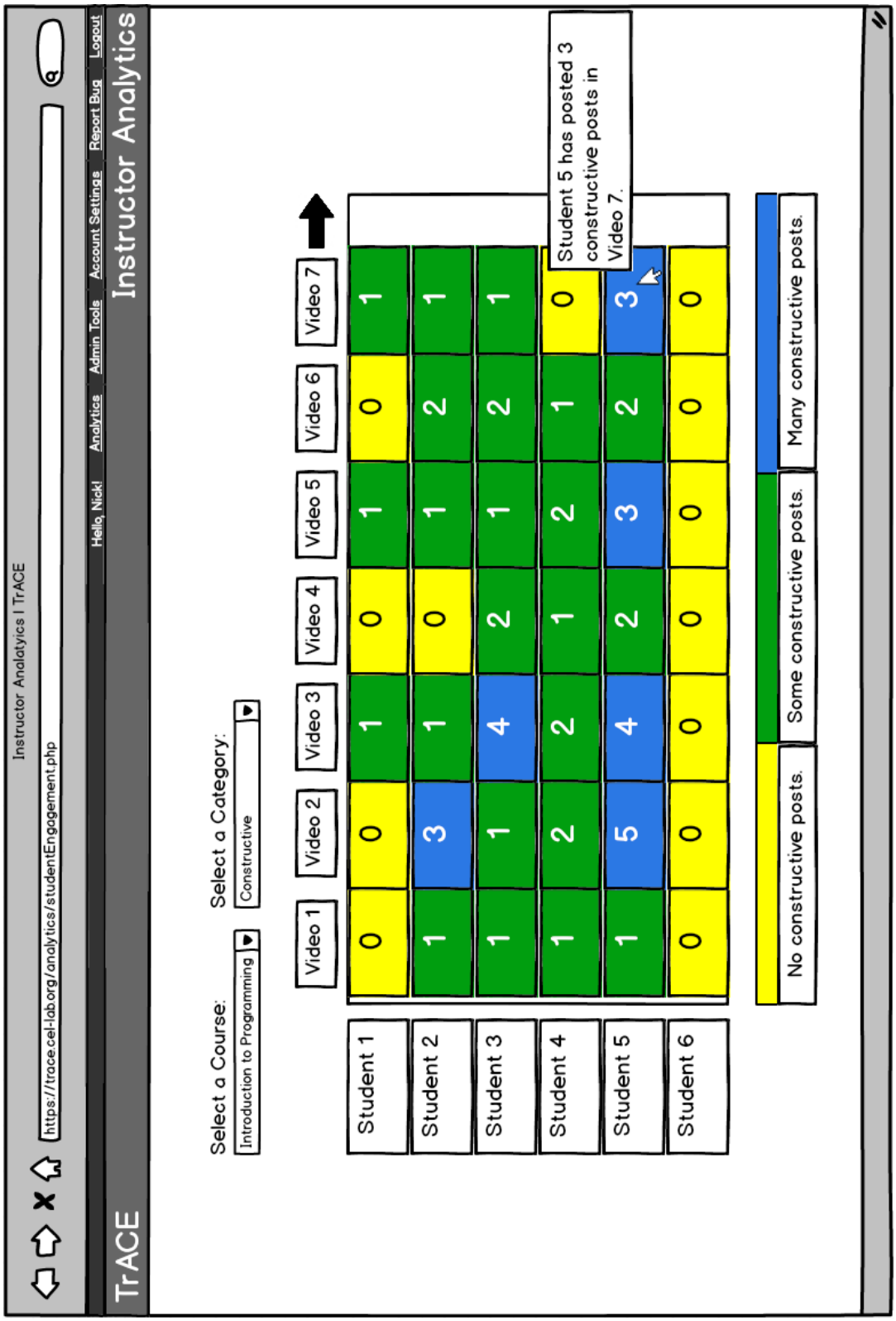

FIGURE 5.3: A chart displaying the number of annotations a each student has posted in each video in some engagement category.

The last analytic in figure 5.4 provides similar information to that of the previous analytic, but enables instructors to better view how a student's engagement with course content is changing over time. After selecting a student to view, the total number of annotations posted by that student in each engagement category is shown at the top, along with the proportion of annotations that category contains. Underneath is a bar graph that shows, for each video, how many annotations in each category that student has posted. With the example data in the figure, it is shown that student 5 has posted one *constructive* annotation, two *active 1* annotations, and one *active 2* annotations in video 3. As the previous analytic did, this one also helps with the categories "Experiencing Difficulty" and "Individual Student Intervention."

A use-case that this analytic would be better for than the previous analytic is that of discovering students who are becoming more disengaged as the course proceeds. Where the previous analytic is better at discovering consistently disengaged students, this one can be used to more quickly identify students who were previously very engaged, but are experiencing a drop-off in engagement, possibly warranting individual student intervention would be helpful for that student. The post counts and proportions displayed above the graph in the analytic help with this task, as the instructor can compare a student's average post proportions with the post proportions of the most recent videos. Flipping this scenario, an instructor might use this analytic after concluding that intervention is needed to see if that student's engagement with course content is improving.

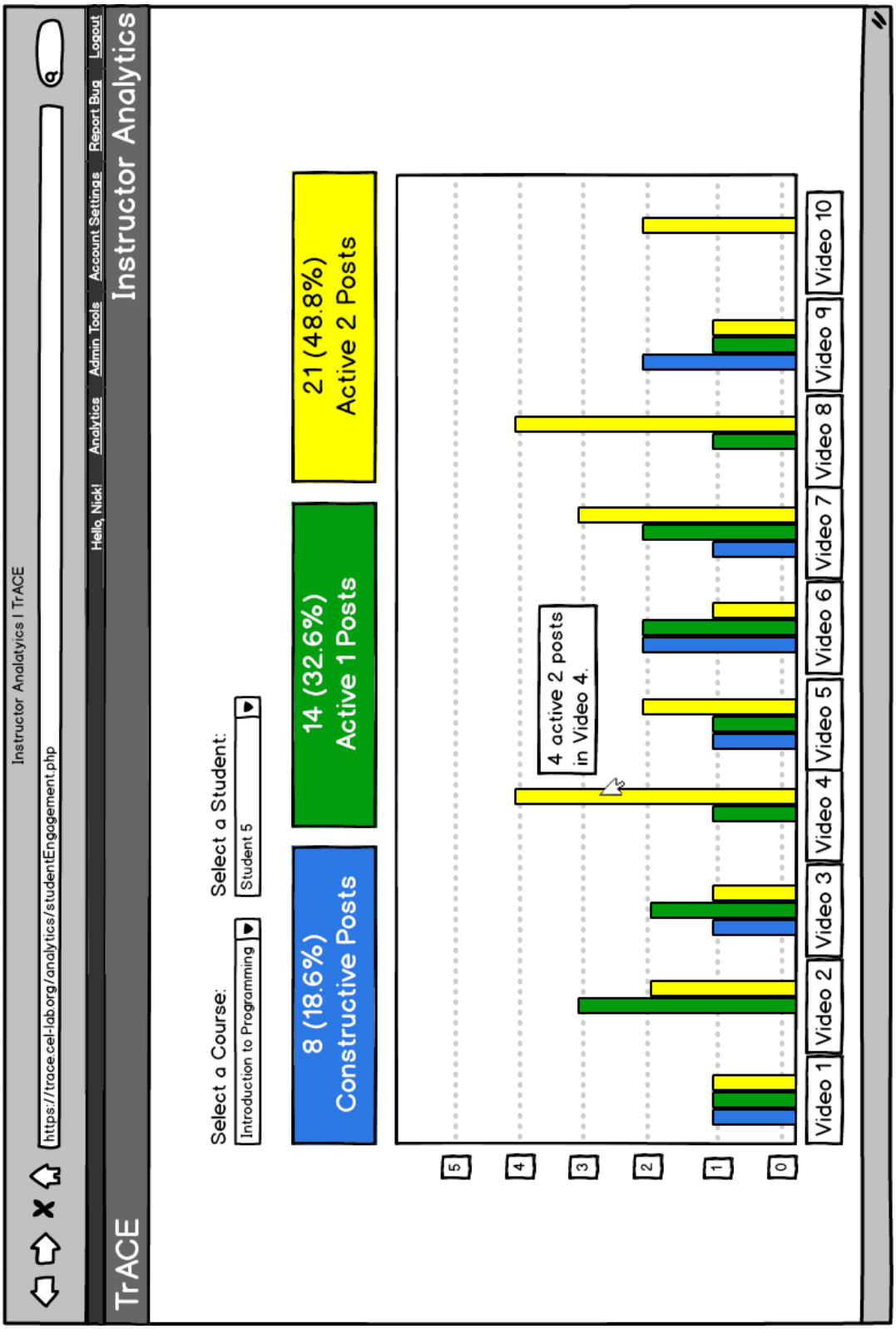

FIGURE 5.4: A mockup of the TrACE video player with added features with engagement data.

5.3 In Video Use

In addition to using engagement data to create analytics for instructors, the data could be used to enhance the ability of instructors to learn about student engagement while going through annotations on the video page. Figure 5.5 shows the TrACE video player, but the data from each annotation being categorized by engagement is used to enhance the instructor's ability to work through annotations. When an instructor opens a video, the colored bars on the right of each annotation represents its engagement level, and annotations can be filtered by engagement level using the drop down at the top. These features would help with all of the categories mentioned from the thematic analysis of instructor use of TrACE. As instructors are going through the video annotations, the blue bars representing *constructive* posts would help them discover where important or detailed discussion might be happening within the video, and the ability to filter annotations by engagement would streamline this further. These features would help instructors get a better idea of both what students individually are doing and what the class as a whole is doing.

The ability for the video player page to filter annotations by engagement could also be tied in with the previous analytic mockups to make them more useful. For example, by clicking on the dot representing 20 constructive posts in video 6 in figure 5.1, the user could be sent to the video page for video 6 with annotations pre-filtered by constructive posts. Annotations could also be filtered by student, so an analytic such as in figure 5.3 could send the user to video 7 with annotations filtered to constructive annotations posted by student 5 if they click on the box for student 5 and video 7. These links to the video pages would more easily allow an instructor to dig deeper into the engagement of their students after they discover interesting behaviors in the analytics.

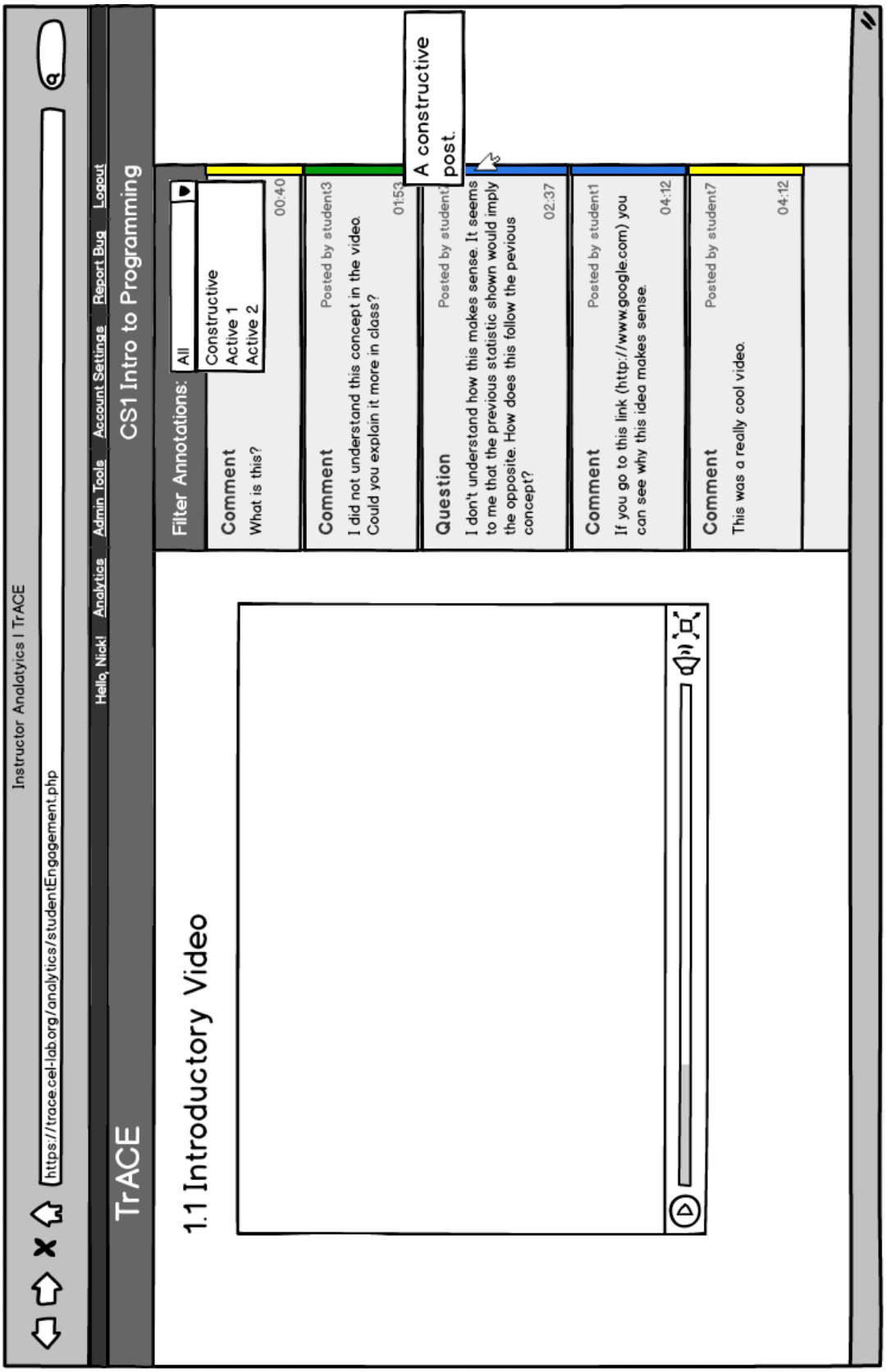

FIGURE 5.5: A bar graph displaying the number of annotations in each engagement category a student has posted in each video.

5.4 Summary and Design Limitations

Presented in this chapter have been four learning analytic mockups, two that enable instructors to observe class trends in engagement and two that allow instructors to observe trends in student behavior, as well as what adding engagement features to a video player might look like. These analytics aim to resolve some of the issues with online learning brought up in the introduction by making instructors more aware of how students are engaging with course content and better enabling instructor intervention. Furthermore, these analytics all address specific pieces of information that instructors reported wanting in the analytics during interviews about their usage of TrACE.

It is important to note that this chapter is exploratory, and not meant to be an exhaustive list of how a system categorizing engagement can be used in learning environments. Additionally, these designs have revolved around the system studied in this thesis, TrACE. Being grounded in previous research on online learning environments, these designs are likely a solid starting point for building within other lecture video systems, but the context and affordances of those systems may necessitate changes. The specific affordances of those systems may also be able to enhance these analytic designs. For example, although these mockups have been designed with only the machine learning model from this thesis in mind, TrACE has the ability to embed quiz questions and forced pauses into videos. Results from these types of interactions could be combined with student engagement data to generate analytics that provide a higher overview of engagement.

One limitation of these mockups is their use of the three engagement category names, *constructive*, *active 1*, and *active 2*. Instructors will likely not know what these mean and therefore not understand these metrics having not worked with the ICAP framework before. Before actually building these analytics, replacement names for the three categories would

be needed, and explanations of what each category stands for would need to be provided to instructors for them to fully understand them.

Another limitation is that these metrics are all based on a machine learning model, and even if accuracy can be significantly improved from the results in the previous chapter of this thesis, these metrics can not be taken as absolute fact about student engagement or as absolutely representing their understanding of course material. This is the case with any data based on machine learning, as 100% accuracy can not be guaranteed and should not be expected.

As discussed in the introduction, the goal of testing this application of machine learning and designing these analytics is to help instructors build a better understanding of their students' engagement while using online learning systems, especially when there are hundreds of students and instructors can not possibly read the posts of every student. The analytics in this section aim to point instructors in the right direction, and filter information when going through every data point is not possible. The instructor should then further explore the system themselves, hopefully using the links to the video pages from the analytics to make a more personal decision of whether intervention would be helpful for a student, or to build a broader understanding of how the class is engaging with the material as a whole. This is of particular concern as instructors often use TrACE analytics for grading [18], and anecdotally, instructors may become too heavily reliant on the analytics when they are short for time. These mockups have been designed with exploration in mind to prevent these use cases from occurring.

Chapter 6

Conclusion

The increasing use of online systems for learning has given birth to new fields of research looking at factors impacting student learning in the online context. This thesis has attempted to address one of these factors, that of increasing instructor awareness of student engagement during the learning process in online systems. Previous research in this area has mostly involved analyzing the log data stored in these systems, but only few recent studies have begun to apply natural language processing methods to the text artifacts generated by students. Additionally, these prior studies have largely only studied student behavior or created models for performance prediction, and have not taken steps to use this research to actually impact student success. This thesis has used these previous studies, the ICAP framework, and previous research with the ICAP framework to provide a machine learning model and set of analytic mockups that future online learning systems can potentially use to impact student success.

This thesis has provided three main contributions. First, a machine learning method that is capable of categorizing student posts by their level of engagement has been described in this work. The performance evaluation suggests that this method is indeed feasible, and

future work with larger data sets can likely improve the accuracy of this method. Second, statistical analyses and qualitative discussions of the results from those analyses have shown what features are important for machine learning, and, for the bag-of-words features, that course keywords are the features important for engagement. Future iterations of similar engagement models should be more capable of generalizing these methods to many courses using the insights provided by these analyses. Lastly, mockups of what learning analytics might be built on top of this engagement classification method have been designed. These designs were informed by the problems with online systems identified by previous research and a prior study performing a thematic analysis on interviews and journals made by instructors using TrACE. Although the mockups were based in the TrACE environment, they should be applicable to other learning systems that afford similar discussion and posting functions.

6.1 Limitations and Future Work

One limitation of this work is the relatively small sample size of annotations, and the small number of courses that had a large enough sample size to be included in this thesis. The especially small sample of *constructive* annotations in the data sets seems to have affected the performance of identifying *constructive* annotations, and performance evaluation on the information assurance data set was not possible due to the very small sample of annotations. Only three courses were offered in enough semesters and generated enough annotation data to be useful for analysis. Unfortunately, these three courses were all STEM, and specifically information technology related classes. Being able to analyze data from courses in different fields would have been useful in further studying how generalizable these methods are. It may be that features have different correlations with engagement in different areas of study,

and these methods are not useful outside of STEM courses. Future work should try to obtain data from various fields of study to see if this is true.

Another limitation is that of replicating these methods in other systems. This thesis was not able to conclude that these methods are generalizable to many courses, but the method of discourse in TrACE (annotations tied to a time and place on the video) may affect the length or depth of language used. Other systems may have to go through the process of manually annotating data to train a model for their system. On the other hand, the insights and correlations that features have with engagement should be the same and still be able to inform the development of engagement classification methods in different systems. The analytic designs especially should still be applicable to other systems, as they were informed by prior literature working with many different types of learning and instructors using TrACE who teach in both the flipped and online formats.

Looking at the performance evaluation portion of this thesis, only one version of a neural network, a multilayer perceptron, was evaluated. This method was chosen based on the prevalence of neural networks used in previous research conducting various natural language tasks. Future research that looks to implement such a system should spend more time testing different machine learning algorithms to improve accuracy as much as possible. The scope of this thesis was to test the feasibility of this application of machine learning, so it did not look into what kind of biases the multilayer perceptron may have and how that affected performance. Similarly, this thesis did not look into what exactly was causing the lower recall metrics for constructive posts in the performance evaluation. Although outside the scope of this thesis, taking a qualitative look into what kinds of posts were being classified incorrectly may be able to shed light on what features could be added or removed to increase performance.

This thesis has laid the groundwork for the analytics to be built on top of engagement categorization data, but there is still work left to be done for systems planning on implementing this design. As previously discussed, this thesis could only look at how general these methods were to a certain extent, and systems with more varied courses and data should continue to research this. Likely, there are other natural language features that correlate with engagement, and future work should look at testing more complex features and larger numbers of them to see how the accuracy of this model can be improved. Additionally, the hypothesis presented in this thesis that bag-of-words features can be replaced by one or multiple course keyword features should be tested. Finally, after actually building the system or prototypes of it, conducting usability tests as well as studies using other qualitative methods to see how analytic designs can be improved should be conducted. Although the designs in this work took inspiration from many different sources, there may be other use cases or desirable features that were missed.

6.2 Final Thoughts

This thesis has attempted to use the large amounts of data generated in TrACE to explore the feasibility of a system that would use machine learning to impact instructors use of the system, and therefor student success. As discussed earlier, machine learning applications and studies in the domain of online learning are plentiful, but far too often do they simply look at predicting performance or behaviors without actually implementing anything that impacts users. Hopefully the grounding in educational psychology, consideration of how instructors use TrACE analytics while designing mockups, and focus on an issue that instructors themselves are concerned about brings this work closer to having an impact on online education.

Bibliography

[1] Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., and McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 114–123. ACM Press.

[2] Arnold, K. E. and Pistilli, M. D. (2012). Course signals at purdue: using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270. ACM.

[3] Bergmann, J. and Sams, A. (2012). *Flip your classroom: reach every student in every class every day.* International Society for Technology in Education.

[4] Bransford, J., B. A. C. R. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition.* The National Academies Press, Washington, DC.

[5] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, 8(1):13–25.

[6] Brinton, C. and Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2299–2307.

[7] Chi, M. T. H. (2009). Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science*, 1(1):73–105.

[8] Chi, M. T. H. and Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4):219–243.

[9] Claudia Mazziotti, Alice Hansen, B. G. (2016). It ain't what you do, it's the way that you do it: Investigating the effect of students' active and constructive interactions with fractions representations. *12th International Conference of the Learning Sciences*, pages 753–760.

[10] Coleman, C. A., Seaton, D. T., and Chuang, I. (2015). Probabilistic Use Cases: Discovering Behavioral Patterns for Predicting Certification. In *Proceedings of the Second ACM Conference on Learning @ Scale*, pages 141–148. ACM Press.

[11] Craig, S., Graesser, A., Sullins, J., and Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250.

[12] Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., and Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 6–14. ACM.

[13] Dazo, S. L., Stepanek, N. R., Fulkerson, R., and Dorn, B. (2016). An Empirical Analysis of Video Viewing Behaviors in Flipped CS1 Courses. In *21st Annual Conference on Innovation and Technology in Computer Science Education*, pages 106–111. ACM Press.

[14] Djuriš, J., Medarević, D., Krstić, M., Vasiljević, I., Mašić, I., and Ibrić, S. (2012). Design Space Approach in Optimization of Fluid Bed Granulation and Tablets Compression Process. *The Scientific World Journal*, 2012:1–10.

[15] d'Oliveira, C., Carson, S., James, K., and Lazarus, J. (2010). MIT OpenCourseWare: Unlocking Knowledge, Empowering Minds. *Science*, 329(5991):525–526.

[16] Dorn, B., Schroeder, L. B., and Stankiewicz, A. (2015). Piloting TrACE: Exploring Spatiotemporal Anchored Collaboration in Asynchronous Learning. In *Proceedings of the 18th ACM Conference on Computer Suppotred Cooperative Work and Social Computing*, pages 393–403. ACM Press.

[17] D'Mello, S., Lehman, B., Pekrun, R., and Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170.

[18] Elson, J. S. (2016). Formative assessment in an online asynchronous learning environment. Master's thesis, University of Nebraska at Omaha.

[19] Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6):304.

[20] Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. Wiley.

[21] Garrison, D. R., Anderson, T., and Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, 2(2):87–105.

[22] Guo, P. J., Kim, J., and Rubin, R. (2014a). How video production affects student engagement: an empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale*, pages 41–50. ACM Press.

[23] Guo, P. J., Kim, J., and Rubin, R. (2014b). Understanding In-Video Dropouts and Interaction Peaks in Online Lecture Videos. In *Proceedings of the First ACM Conference on Learning @ Scale*, pages 41–50. ACM Press.

[24] Herreid, C. F. and Schiller, N. A. (2013). Case studies and the flipped classroom. *Journal of College Science Teaching*, 42(5):62–66.

[25] Hill, P. (2012). Online educational delivery models: A descriptive view. EDUCAUSE.

[26] Jayaprakash, S. M., Lauría, E. J. M., Gandhi, P., and Mendhe, D. (2016). Benchmarking student performance and engagement in an early alert predictive system using interactive radar charts. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 526–527. ACM Press.

[27] Jordan, K. (2016). MOOC completion rates. http://www.katyjordan.com/.

[28] K. E. Arnold, Z. Tanes, A. S. K. (2010). Administrative perceptions of data-mining software signals: Promoting student success and retention. In *The Journal of Academic Administration in Higher Education*, pages 29–39.

[29] Kizilcec, R. F., Piech, C., and Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the*

Third International Conference on Learning Analytics and Knowledge, pages 170–179. ACM.

[30] Kort, B., Reilly, R., and Picard, R. W. (2001a). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, page 43. IEEE.

[31] Kort, B., Reilly, R., and Picard, R. W. (2001b). External representation of learning process and domain knowledge: Affective state as a determinate of its structure and function. In *Workshop on Artificial Intelligence in Education*.

[32] Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., and Siemens, G. (2016). Towards automated content analysis of discussion transcripts: a cognitive presence case. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 15–24. ACM Press.

[33] Lage, M. J., Platt, G. J., and Treglia, M. (2000). Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment. *The Journal of Economic Education*, 31(1):30.

[34] Lee, D. M. C., Rodrigo, M. M. T., d Baker, R. S., Sugay, J. O., and Coronel, A. (2011). Exploring the relationship between novice programmer confusion and achievement. In *International Conference on Affective Computing and Intelligent Interaction*, pages 175–184. Springer.

[35] Liu, Z., Pataranutaporn, V., Ocumpaugh, J., and Baker, R. (2013). Sequences of frustration and confusion, and learning. In *Educational Data Mining 2013*.

[36] Marzouk, Z. and Winne, M. R. P. (2016). Generating Learning Analytics to Improve Learners' Metacognitive Skills Using nStudy Trace Data and the ICAP Framework. *Learning Analytics for Learners Workshop*.

[37] Menekse, M., Stump, G. S., Krause, S., and Chi, M. T. H. (2013). Differentiated Overt Learning Activities for Effective Instruction in Engineering Classrooms: Differentiated Overt Learning Activities. *Journal of Engineering Education*, 102(3):346–374.

[38] Miller, G. A. (1995). Wordnet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

[39] Schroeder, L., Xue, F., and McGivney, R. (2013). Flipping three different mathematics courses: Common conclusions and plans for the future. *MathAMATYC Educator*, 4(2):63–67.

[40] Senn, G. J. (2008). Comparison of face-to-face and hybrid delivery of a course that requires technology skills development. *Journal of Information Technology Education*, 7(20):267–284.

[41] Severance, C. (2012). Teaching the world: Daphne koller and coursera. *Computer*, 45(8):8–9.

[42] Shafiq, M. Z., Ilyas, M. U., Liu, A. X., and Radha, H. (2013). Identifying leaders and followers in online social networks. *IEEE Journal on Selected Areas in Communications*, 31(9):618–628.

[43] Singh, V., Abdellahi, S., Maher, M. L., and Latulipe, C. (2016). The Video Collaboratory as a Learning Environment. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 352–357. ACM Press.

[44] Smith, E. A. and Senter, R. J. (1967). Automated readability index. Technical report, DTIC Document.

[45] Wang, X., Wen, M., and Rosé, C. P. (2016). Towards triggering higher-order thinking behaviors in MOOCs. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 398–407. ACM Press.

[46] Wang, X., Yang, D., Wen, M., Koedinger, K., and Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society*.

[47] Wen, M., Yang, D., and Rosé, C. P. (2014a). Linguistic Reflections of Student Engagement in Massive Open Online Courses. In *The 8th International AAAI Conference on Weblogs and Social Media*.

[48] Wen, M., Yang, D., and Rosé, C. P. (2014b). Sentiment analysis in MOOC discussion forums: What does it tell us. *Proceedings of Educational Data Mining*, 1.

[49] Wertsch, J. (1985). *Vygotsky and the social formation of mind*. Harvard University Press.

[50] Wolff, A., Zdrahal, Z., Nikolov, A., and Pantucek, M. (2013). Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 145–149. ACM.

[51] Yuan, L. and Powell, S. (2013). MOOCs and open education: Implications for higher education. Technical report, JISC Cetis.

[52] Zhang, D., Zhao, J. L., Zhou, L., and Nunamaker, Jr., J. F. (2004). Can e-learning replace classroom learning? *Communications of the ACM - New architectures for financial services*, 47(5):75–79.

[53] Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y., and Paquette, L. (2016). Longitudinal engagement, performance, and social connectivity: a MOOC case study using exponential random graph models. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 223–230. ACM Press.