To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns

Caitlin Mills[1], Nigel Bosch[2], Art Graesser[3,] and Sidney D'Mello[1,2]

Departments of Psychology[1] and Computer Science[2], University of Notre Dame
Notre Dame, IN 46556, USA
{cmills4|pbosch1|sdmello}@nd.edu

Department of Psychology and Institute for Intelligent Systems[3], University of Memphis,
Memphis, TN 38152, USA
{a-graesser}@memphis.edu

Abstract. This research predicted behavioral disengagement using quitting behaviors while learning from instructional texts. Supervised machine learning algorithms were used to predict if students would quit an upcoming text by analyzing reading behaviors observed in previous texts. Behavioral disengagement (quitting) at any point during the text was predicted with an accuracy of 76.5% (48% above chance), before students even began engaging with the text. We also predicted if a student would quit reading on the first page of a text or continue reading past the first page with an accuracy of 88.5% (29% above chance), as well as if students would quit sometime after the first page with an accuracy of 81.4% (51% greater than chance). Both *actual* quits and *predicted* quits were significantly related to learning, which provides some evidence for the predictive validity of our model. Implications and future work related to ITSs are also discussed.

Keywords: engagement, disengagement, affect detection, reading, ITSs

1 Introduction

One of the benefits afforded by intelligent tutoring systems (ITSs) and other advanced learning technologies is the students' ability to move at their own pace through learning sessions. In many systems, students have choice over the topics and activities they engage in. Importantly, they can also choose how long to spend on each one. However, one caveat to this type of choice is that disengagement can occur before activities or topics are completed, leaving vital information unseen. Therefore, identifying when disengagement will occur may help inform timely interventions, such as temporarily suppressing choice or providing motivational messages to persist [1], as well as development of educational materials that keep students engaged in order to achieve learning goals.

There has been a growing interest in automatically detecting students' affective states and engagement during learning (see [2] for a review). One focus, in particular,

has been on identifying behaviors associated with engagement/disengagement during learning because of the necessity of engagement for learning [3]. In fact, previous research has had success in modeling and detecting various types of disengaged behaviors within ITSs [4–9]. For example, an automatic detector for "gaming" the system can reliably detect when students exploit the system to get correct answers [4]. Another detector also made it possible to recognize if a student is purely off-task or engaging in on-task conversation [10]. These types of detectors have led to helpful design interventions, as well as more accurate student models of learning [11].

Previous work has also been able to classify different levels of engagement using log files. Cocea and Weibelzahl [12] classified 10-minute intervals of a learning session as one of three levels of engagement: engaged, disengaged, or neutral. Ground truth was achieved from labels provided by expert human coders. This study reported accuracies 71% greater than baseline (Cohen's kappa = .713) using features extracted from log file information, such as reading behaviors (i.e., average time, number of pages) and test information (i.e., average time, number of tests, correct answers). This model displayed impressive accuracies for diagnosing students' current level of engagement during the specified 10-minute intervals and appears to generalize across multiple learning environments [13]; however, predictors of future engagement have not yet been established.

All of the detectors mentioned thus far have focused on a specific aspect of engagement (or disengagement), such as behaviors like gaming the system. Indeed, engagement has been operationalized in numerous ways due to its multi-faceted nature [14]. Specifically, engagement can be thought of as encompassing three distinct components: (a) affect (e.g., positive and negative feelings), (b) behavior (e.g., persistence, effort), and (c) cognition (e.g., goals, self-regulated behaviors) [15]. Typically, disengagement detectors target some combination of these three components, initially relying on external coders to make some inference about the cognitive/affective components based on student behavior or self-report measures. One problem then, as noted by Baker and Rossi [14], is that models of engagement are difficult to validate beyond face validity because engagement is complex, and ground truth is achieved via human judgments, which are inherently subjective.

The current research focus is on a behavioral indicator of disengagement. Specifically, we build a predictor of behavioral disengagement, which we operationally define as the point at which a student opts to stop interacting with (quits) a given activity within a learning session. Importantly, this operationalization of behavioral engagement does not require any external human coders to initially establish ground truth. A distinguishing aspect of this work is that our model is *predictive* in that disengagement on the current activity N is predicted from interaction patterns observed during the previous N-1 activities instead of *diagnostic*, where actions in N are used to detect disengagement in N after it occurs. A predictive model can ostensibly be used to prevent the onset of disengagement, which is advantageous since disengagement and boredom are long-lasting persistent negative states [16].

The instructional reading task in the present research is a self-paced learning task where students control the pace and time spent on each text. Self-paced reading is an important component within a number of interactive learning technologies and ITSs,

such as in Operation ARA!, iSTART, and ELM-ART [17–19]. For example, in Operation ARA!, students read an electronic textbook before engaging in tutorial dialogs. We use sensor-free information from previous activities (i.e., log files of reading patterns) to predict quitting before the current text ever begins. The ability to unobtrusively predict when quitting behaviors will occur provides the foundation for effective design of interventions to keep students engaged.

2 Methods

2.1 Data Collection

Participants. Data was obtained from 173 undergraduate students from a private university in the Midwest and a large public Mid-south university in the US who participated for course credit.

Texts. Students spent a total of 30 minutes completing reading from instructional texts. The reading task consisted of eight texts on scientific research methods topics (disguised measures, gathering data, hypotheses, scientific method, construct validity, variables, criterion of precision, expectancy bias) adapted from a popular textbook [20]. Texts had an average length of 1068 words (*SD* = 35.7) with a Flesch-Kincaid Grade Level score of about 13.5, which is indicative of some difficulty. Order of topics was counterbalanced across students.

Procedure. Students completed an informed consent and a short trial to familiarize themselves with the interface. Each student was then left alone in a small room for 30 minutes with the reading interface. No other devices or distractions, such as a watch or cell phone were permitted. Students were presented with a blank screen with a button labeled READ to begin the reading task. A text was presented once a student selected the READ button. Texts were presented one page at a time with 77 words per page. Students could use the right and left arrow keys to navigate through the text with the ability to move backward to previous pages or forward to the next page. Students had the capability of quitting the text at any point in time by pressing the 'C' key ("Change to a different text"). If students hit the 'C' key, a new text would appear. Students could press the 'C' key up to seven times and receive a new text (eight texts). Only data from the first time students viewed each text were analyzed in order to avoid familiarity biases after seeing a text multiple times. In sum, over the course of the 30 minutes, students were able to read as much or as little of each text as they chose.

As a learning measure, students completed a posttest involving 48 multiple-choice questions (six per text) about the information from the eight texts after the reading session. Questions were developed in adherence to the Graesser-Person question-asking taxonomy [21]. The questions targeted specific sections in the text, such that answers were not apparent unless the targeted section of text was read.

Quitting Behaviors. Students' reading time information (e.g., how long they spent on each page) was collected during the reading task. Every text was classified as *Quit*, *Completed*, or *Timeout* based on how the student interacted with the text. Instances labeled Quit consisted of texts that students started reading, but hit the 'C' key to exit

the text before reaching the end of the text. Completed instances were texts that were read by students in their entirety. Finally, an instance was labeled as Timeout if the learning session was interrupted in the midst of reading due to the 30 minute time limit, and therefore could not be classified as Quit or Completed. The instances (texts) that were labeled as Timeout were removed from the dataset because we were not interested in a forced exit from a text. In total, there were 911 instances used to build models, where students either quit (n = 311) or completed (n = 600) a text after beginning to read it for the first time, thereby yielding a 34% rate of quitting. On average, students quit texts after reading 32.9% of the pages (SD = 28.3%).

2.2 Model Building

Feature Engineering and Selection. A total of 18 features were computed from reading behaviors and reading times. For each text analyzed (text N), two types of features were extracted: previous text information (text N-1) and cumulative previous texts information [e.g., features from all previous texts (1 to N-1) averaged]. No feature used any information from the current text being classified or any text that was viewed later, which is essential for predictive modeling. Table 1 contains a list of the features that were computed based on the logged reading behaviors (e.g., reading times, quit behaviors).

Using a backward feature selection method, features from the *previous text* feature set were removed one at a time depending on model performance after removing a feature[1]. If model performance declined, the feature was retained for the final model. Next, features from the *cumulative previous texts* feature set were removed in the same manner. Finally, backward selection was used on the combined set of remaining features from the two feature sets to produce a final set of features for each classification task. There were no features that correlated higher than .80 or higher, which was used as a threshold to remove correlated features.

Supervised Classification. We used supervised machine learning to build predictors for three different classification tasks. The first task attempted to classify if a student would quit at any point during a particular text vs. completely read the text. The second task attempted to classify if a student would quit on the first page of the text vs. continue reading past the first page (even if they might eventually quit at some point). Finally, the third task aimed to classify if a student would quit at any point past page one vs. completely read the text. Six binary classification algorithms provided in Rapid Miner were used for each of the models, including Bayes Net, RIPPER (JRip implementation), C4.5 (J48 implementation), Naïve Bayes, SMO, and VFI.

Model Validation. All models were evaluated using leave-one-student-out cross-validation, in which k-1 students are used in the training data set. The model is then tested on the student who was not used in the training data. This process is repeated until every student has been used as the testing set one time. The average results from

[1] We also tested models using all 18 features, which exhibited worse performance (assessed via Cohen's Kappa) than each of the three final models using the selected features.

the k iterations provide an estimation of classification accuracy. Cross-validating at the student level increases confidence that models will be more generalizable when applied to new students because the testing and training sets are independent.

Table 1. Description of features and indication of which final model(s) each was included (+).

Features	Quitting on Any Page vs. Completing	Quitting on Page 1 vs. Continuing	Quitting After Page 1 vs. Completing
Previous Text Only			
Page 1 Reading Time (RT)		+	+
Quit On Page 1 (Yes/No)		+	+
Location of Quit (First 3 Pages, After 3 Pages, None)		+	+
Max Page Number Seen			
Median Page Reading Time (RT)			
Minimum Page Reading Time			
Maximum Page Reading Time			
Proportion of Text Seen		+	+
Reading Time 1 Page Before Exit	+	+	
Proportion of Pages < 5s Reading Time	+	+	+
Total Reading Time		+	+
Text Exit (Quit/Completed)	+		+
Cumulative Previous Texts Seen			
Maximum Page Number Seen	+	+	
Median Page Reading Time		+	
Minimum Page Reading Time			
Maximum Page Reading Time			
Proportion of Pages < 5s Reading Time	+	+	
Total Reading Time			

Metrics. Classification accuracy was evaluated using precision, recall [22], and Cohen's kappa [23]. Precision represents the percentage of texts classified as Quit that were actually Quit. Recall represents the percentage of texts that were actually Quit and also correctly classified as Quit. Cohen's kappa takes base rates into consideration and indicates the degree to which the model is better than chance (kappa of 0) at correctly predicting whether the text will be Quit or Completed. A kappa value of -0.5 or 0.5 would indicate the model is classifying -50% worse or 50% better than chance, respectively. We also report percent correctly classified (accuracy), but caution that this should be interpreted cautiously since class imbalance tends to inflate accuracy.

3 Results and Discussion

3.1 Quitting on Any Page vs. Completing the Text

The first classification was to attempt to predict whether a student would quit a text at any point or complete the text. The six classifiers were used to predict quitting based on the features extracted from text(s) previously presented to the student (see above). The best model for predicting overall quitting behavior used the Bayes Net algorithm. The kappa for this model indicates the model's performance is 48.4% higher than chance. Five features were used in this best model (indicated in Table 1). Model fit statistics are presented in Table 2.

Table 2. Performance measures for the three classification tasks

	Quit Class		**Completed/ Continued Class**			
	Precision	**Recall**	**Precision**	**Recall**	**Kappa**	**Accuracy**
Any Page	64.8%	68.2%	83.1%	80.8%	.484	76.5%
First Page	38.7%	33.0%	92.9%	94.4%	.293	88.5%
Subsequent Pages	67.5%	60.5%	85.9%	89.2%	.514	81.4%

We also examined the confusion matrix for this predictor (Table 3). It is notable that both true positives and true positives were higher than false positives or false negatives. Given a prediction of Quit, odds were nearly 2:1 (64.8% precision) that the prediction is correct (a "hit" rather than a "false alarm"), and so an intervention can be given with a good degree of confidence.

3.2 Quitting on the First Page vs. Continuing

The next classification task attempted to predict if students would quit on the first page vs. continue reading, which occurred 10% of the time. Predicting these instances may provide information for more immediate interventions before quitting occurs on page one. For this task, Quit labels were restricted to the cases where students quit the text on the first page. Any quit past page one is classified as a *Continue Past Page One*. The best classifier was a Bayes Net algorithm using 10 features (see Table 1). Performance measures are provided in Tables 2 and 3, respectively.

This model was able to classify texts where students quit on the first page 29.3% higher than chance using information from previous text(s). Although this predictor does not perform as well as the previous model, this model provides an important classification at a relatively small window size (page level). The confusion matrix for

the first page Quit model illustrates the class imbalance well. Due to the large proportion of Continue Past Page One instances (.903), Quit instances were not likely to be detected as well as Quit instances on any page. Interventions given based on these predictions must be especially cautious, using a "fail soft" approach. The low precision (38.7%) implies that less than half of the Quit predictions will be correct, due largely to the class imbalance.

Table 3. Confusion matrices for the three classification tasks.

Any Page	**Predicted Quit**	**Predicted Completed**	**Priors**
Actual Quit	0.68 (hit)	0.32 (miss)	0.34
Actual Completed	0.19 (false alarm)	0.81 (correct rejection)	0.66
First Page	**Predicted Quit**	**Predicted Continued**	**Priors**
Actual Quit	0.33 (hit)	0.67 (miss)	0.10
Actual Continued	0.06 (false alarm)	0.94 (correct rejection)	0.90
Subsequent Pages	**Predicted Quit**	**Predicted Completed**	**Priors**
Actual Quit	0.61 (hit)	0.39 (miss)	0.27
Actual Completed	0.11 (false alarm)	0.89 (correct rejection)	0.73

3.3 Quitting After the First Page vs. Completing the Text

The third classification task attempted to predict quitting once students read past the first page vs. completing. Since 10% of texts were quit on page one, it is also useful to understand when students will quit after reading past the initial first page. Classifying quitting once students read past page one will allow interventions to target students who are moving through the text (past the initial page), yet decide to stop before completing the entire text.

The cases where students quit on the first page were not included in this task, leaving 223 instances labeled as Quit and 600 labeled as Completed. The best classifier was a C4.5 classifier, which was able to perform 51.4% higher than chance (see Tables 2 and 3 for performance summary). Interestingly, this model differed from the first two classifications tasks, as only the features containing information from the previous text were included in the model (see Table 1). Precision for this model was 67.5% and had a lower proportion of false alarms than in the "Any Page" model, indicating some potential for use with interventions.

3.4 Predictive Validity

We also examined the relationship between posttest performance and quitting. First, we correlated students' proportion of correct responses on the posttest (posttest per-

formance) with their proportion of actual quits, Pearson's r = -.314, p < .001. Indeed, this negative correlation provides some validation for the use of quitting as a measure of behavioral disengagement, as disengagement is associated with negative learning [5].

It is also important to establish whether posttest performance was related to our model's predicted quits. Students' posttest performance was also correlated with the proportion of predicted quits, based on model classification (i.e., Quit vs. Finished using the Bayes Net algorithm), r = -.332, p < .001. This correlation gives us some confidence in our model's predictive validity, since our predicted quits are negatively related to learning as well.

Finally, we also investigated the relationship between actual quits and predicted quits at the student level. The proportion of students' actual quits was highly correlated with the proportion of predicted quits (as predicted using the Bayes Net algorithm), r = .934, p < .001. This positive relationship gives us further confidence in our predictor, as students' quitting behavior was closely tied to the model's predictions.

4 General Discussion

We developed three models of quitting by analyzing log files from previous texts: (1) any point during a text vs. completing the text (kappa of .484), (2) on the first page vs. continuing reading (kappa of .293), and (3) past the first page vs. reading to completion (kappa of .514). Importantly, we are attempting to predict future behavior before the activity is even started from easily available reading measures, so this form of modest kappa is expected. Additionally, the kappa values achieved using these predictors are similar to those found in previous disengagement detectors [24, 9], however meaningful comparisons of results are complicated by differences in how disengagement is conceptualized.

The features that were used in the final models reveal that reading times on key pages are important for predicting quitting. For example, reading time on the page immediately before quitting the previous text was included in two of the final models and the proportion of pages with reading time less than five seconds was included as a feature in all three final models. Furthermore, the reading time on the first page was included in two out of three final models. Previous quitting behavior was also relevant in these predictors. In fact, students previously quitting on the first page, as well as what section of the text they quit (first three pages, after first three pages, or completed) were also relevant features in two of the final models. These predictors indicate that past (reading) behavior can be a good indicator of future behavior.

Predicting quitting behaviors may open up new avenues for interventions and instructional designs in order to facilitate better learning. When disengagement behaviors, such as gaming the system, are detected, a system can reactively respond by reintroducing the content that is missed due to gaming for improved learning [11]. The predictors presented in this paper are an initial step for interventions that can occur *proactively*, since the prediction is made before the text is even read. For example, the utility of the text can be highlighted as a potential motivator to continue if

quitting is predicted [25]. Or the system might suggest a change of topics or that the student may take a short break.

It is important to note that these models are not without limitations. First, these models were fit using an instructional reading task, which may not generalize to other learning environments. Second, our results cannot be generalized beyond the current sample. Third, since this study was conducted in the lab, future work should investigate the effectiveness of similar models using log files from actual ITS learning sessions. Future work should also include attempts to combine these reading behavior features with other trait-based features, such as prior knowledge and interest, which might further improve prediction rates. This paper provides initial groundwork on predicting behavioral disengagement via quitting behaviors, but we believe further development of these types of models are promising for adaptive ITSs to intervene before the moment of disengagement occurs.

Acknowledgment. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

5 References

1. Kelly, K.M., Heffernan, N., D'Mello, S., Namais, J., Strain, A.: Added Teacher-Created Motiational Video to an ITS. The Twenty-Sixth International FLAIRS Conference. pp. 503–508. AAAI Press, Menlo Park, CA (2013).
2. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. Affect. Comput. IEEE Trans. On. 1, 18–37 (2010).
3. Pekrun, R., Linnenbrink-Garcia, L.: Academic emotions and student engagement. Handbook of Research on Student Engagement. pp. 259–282. Springer (2012).
4. Baker, R.S.J., Corbett, A.T., Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. Intelligent Tutoring Systems. pp. 54–76 (2004).
5. Beck, J.E.: Using response times to model student disengagement. Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments. pp. 13–20 (2004).
6. D'Mello, S., Cobian, J., Hunter, M.: Automatic Gaze-Based Detection of Mind Wandering during Reading. Proceedings of the 6th International Conference on Educational Data Mining. pp. 364–365. International Educational Data Mining Society (2013).
7. Forbes-Riley, K., Litman, D.: When does disengagement correlate with learning in spoken dialog computer tutoring? Artificial Intelligence in Education. pp. 81–89 (2011).
8. Rowe, J.P., McQuiggan, S.W., Robison, J.L., Lester, J.C.: Off-Task Behavior in Narrative-Centered Learning Environments. AIED. pp. 99–106 (2009).

9. Wixon, M., Baker, R.S.J., Gobert, J.D., Ocumpaugh, J., Bachmann, M.: WTF? detecting students who are conducting inquiry without thinking fastidiously. User Modeling, Adaptation, and Personalization. pp. 286–296. Springer (2012).
10. Baker, R.S.J.: Modeling and understanding students' off-task behavior in intelligent tutoring systems. Proceedings of the SIGCHI conference on Human factors in Computing Systems. pp. 1059–1068 (2007).
11. Baker, R.S.J., Corbett, A., Koedinger, K., Evenson, S., Roll, I., Wagner, A., Naim, M., Raspat, J., Baker, D., Beck, J.: Adapting to when students game an intelligent tutoring system. Intelligent Tutoring Systems. pp. 392–401 (2006).
12. Cocea, M., Weibelzahl, S.: Eliciting motivation knowledge from log files towards motivation diagnosis for Adaptive Systems. User Modeling 2007. pp. 197–206. Springer (2007).
13. Cocea, M., Weibelzahl, S.: Disengagement Detection in Online Learning: Validation Studies and Perspectives. IEEE Trans. Learn. Technol. 4, 114–124 (2011).
14. Baker, R.S.J., Rossi, L.M.: Assessing the Disengaged Behaviors of Learners. Des. Recomm. Intell. Tutoring Syst. 155 (2013).
15. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: Potential of the concept, state of the evidence. Rev. Educ. Res. 74, 59–109 (2004).
16. D'Mello, S., Graesser, A.C.: The half-life of cognitive-affective states during complex learning. Cogn. Emot. 25, 1299–1308 (2011).
17. Brusilovsky, P., Schwarz, E., Weber, G.: ELM-ART: An intelligent tutoring system on World Wide Web. Intelligent tutoring systems. pp. 261–269 (1996).
18. McNamara, D.S., Levinstein, I.B., Boonthum, C.: iSTART: Interactive strategy training for active reading and thinking. Behav. Res. Methods Instrum. Comput. 36, 222–233 (2004).
19. Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A.C., Halpern, D.: Operation ARIES!: A serious game for teaching scientific inquiry. Serious Games Edutainment Appl. 169–195 (2011).
20. Rosenthal, R., Rosnow, R.L.: Essentials of behavioral analysis: Methods and data analysis. New York: McGraw-Hill (1984).
21. Graesser, A.C., Person, N.K.: Question asking during tutoring. Am. Educ. Res. J. 31, 104–137 (1994).
22. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006).
23. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46 (1960).
24. Baker, R.S.J., De Carvalho, A.: Labeling student behavior faster and more precisely with text replays. Proceedings of the 1st International Conference on Educational Data Mining. pp. 38–47 (2008).
25. Jang, H.: Supporting students' motivation, engagement, and learning during an uninteresting activity. J. Educ. Psychol. 100, 798 (2008).