Forecasting Student Achievement in MOOCs with Natural Language Processing

Carly Robinson
Harvard University
Cambridge, MA
carlyrobinson@g.harvard.edu

Michael Yeomans
Harvard University
Cambridge, MA
yeomans@fas.harvard.edu

Justin Reich
MIT
Cambridge, MA
jreich@mit.edu

Chris Hulleman
University of Virginia
Charlottesvile, VA
csh3f@virginia.edu

Hunter Gehlbach
University of California, Santa Barbara
Santa Barbara, CA
hgehlbach@education.ucsb.edu

ABSTRACT
Student intention and motivation are among the strongest predictors of persistence and completion in Massive Open Online Courses (MOOCs), but these factors are typically measured through fixed-response items that constrain student expression. We use natural language processing techniques to evaluate whether text analysis of open responses questions about motivation and utility value can offer additional capacity to predict persistence and completion over and above information obtained from fixed-response items. Compared to simple benchmarks based on demographics, we find that a machine learning prediction model can learn from unstructured text to predict which students will complete an online course. We show that the model performs well out-of-sample, compared to a standard array of demographics. These results demonstrate the potential for natural language processing to contribute to predicting student success in MOOCs and other forms of open online learning.

Categories and Subject Descriptors
K.3.1 [Computers and Education]: Computer Uses in Education; Distance Learning – Massive Open Online Course, MOOC, Learner Motivation

General Terms
Algorithms, Measurement,

Keywords
MOOCS, Learning Analytics, Motivation

1.INTRODUCTION
In massive open online courses (MOOCs), scholars have predicted achievement-related dimensions such as persistence and completion through tracking log data to predict which students drop-out [2,6,13], and when [12,22]. While generally successful, these lines of inquiry require several days or weeks of activity data to make reliable predictions. This lag may, however, be too long. The probability of dropout is especially high in the first days and weeks of a course [8,13,16], before tracking log based models have sufficient data to make good predictions. Effective early intervention efforts, therefore, may require reliable preliminary predictors.

Additionally, it is not clear how to interpret drop-out prediction from activity logs. This is because the activity logs are also used to determine whether someone has dropped out. So does this kind of model produce a leading indicator of a future drop-out, or a lagging indicator of a recent decision to drop out? This uncertainty constrains the ability of these models to understand the psychology of why people fail to meet their educational goals.

Many MOOC courses collect pre-course survey data on demographics and students' motivations. Research has shown that the strongest predictor of MOOC completion at the outset of a course are students' ratings of whether they intend to complete the course, and demographics have also been useful, to some degree [12,16]. But beyond that, many structured survey items have proven weak predictors of persistence and completion [5].

While intentions are a strong predictor of course completion, it is clear that many people who intend to complete MOOCs do not do so [8,16]. The gap between intention and action is psychologically rich [4], and offers the most promising applications for prediction tools. Students who wish to complete, but whose self-assessments indicate they are unlikely to do so, may be the most receptive for behavioral interventions, compared to students who do not have course completion as a goal. How, then, might we be able to use pre-course survey data to predict their likelihood of falling into that gap?

To answer this question we model course completion in line with Eccles and colleagues' expectancy-value theory [3]. In addition to declaring their intentions, students also described how useful and relevant the course would be to their lives (i.e., the "utility value" of the course). These reasons are important for translating intention into action. Prior empirical work showed that students' perceived utility value of their classes is correlated with their achievement [1, 9] and interventions that increase students' perceived utility value can have a causal effect on student success [9]. So it is possible that

Conference'10, Month 1–2, 2010, City, State, Country.

heterogeneity in students' utility value is also related to the heterogeneity in their course completion rates.

One problem with this empirical approach is that open-ended text is high-dimensional, which reflects the wide range of motivations and goals that students have - especially in a MOOC, where the student body is quite diverse. Still this presents a problem of how to measure utility value quantitatively, when students are writing qualitative statements. In the current research we employ new methods to predict which students follow through on their intentions to earn certification in a MOOC. Specifically, we take advantage of the Natural Language Processing (NLP) toolbox to better understand how students' unstructured text responses in a precourse survey can predict their later course success.

We build upon recent work that takes a similar methodology and argues for using NLP in MOOC analyses [11,17, 21, 23]. Our own research extends this literature in two ways. First, we apply NLP to pre-course data, rather than in-course data, which could open novel approaches to deploying early in-course interventions to help learners. Second, we are using NLP to understand differences between all people who state an intention to complete the course. This ensures that differences in language are not simply reflecting different intentions. Instead, they reflect variation in why students have come upon those intentions, and how students will strive to achieve them.

In this paper we develop an NLP model to parse structured text and predict course completion in a MOOC. Compared to simple benchmarks based on demographics, we find that our machine learning prediction model makes reliable predictions of future student achievement, even when controlling for stated intentions. These results show that unstructured language data can predict student success and suggest new ways to improve student outcomes.

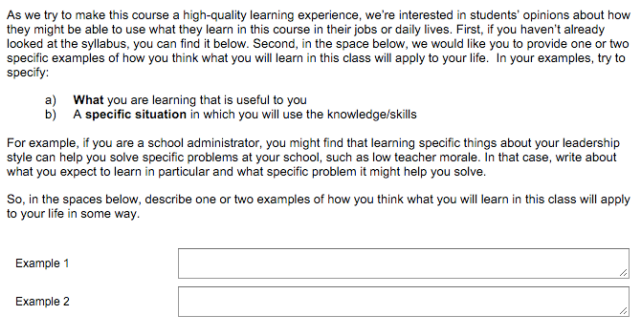
2. METHODS

Data for this study came a HarvardX online course that was education-focused, and included a utility value prompt as part of an experiment that was attempting to improve student persistence In this class (N = 41,946 enrolled students), 47% were female, 25% lived in the United States, and 69% had a Bachelor's degree.

2.1 PROCEDURES

2.1.1 Data collection

The precourse survey covered basic information about the students and the course, including: student intentions and motivations in the course, their prior experience with online learning, and demographic questions. In addition, the pre-course survey also included an open-ended utility value prompt. The prompt was written to elicit the students' expectations of the value they would get from the course (see Figure 1; based on [10]). Specifically, students were asked to name "(a) what you are learning that is useful to you, or (b) A specific situation in which you will use the knowledge/skills." Based on the initial experimental design, this prompt was randomly assigned to half the students in this course as a manipulation [16].


As we try to make this course a high-quality learning experience, we're interested in students' opinions about how they might be able to use what they learn in this course in their jobs or daily lives. First, if you haven't already looked at the syllabus, you can find it below. Second, in the space below, we would like you to provide one or two specific examples of how you think what you will learn in this class will apply to your life. In your examples, try to specify:

a) **What** you are learning that is useful to you
b) A **specific situation** in which you will use the knowledge/skills

For example, if you are a school administrator, you might find that learning specific things about your leadership style can help you solve specific problems at your school, such as low teacher morale. In that case, write about what you expect to learn in particular and what specific problem it might help you solve.

So, in the spaces below, describe one or two examples of how you think what you will learn in this class will apply to your life in some way.

Example 1

Example 2

Figure 1: Text of utility value question

2.1.2 Population filters

We restrict our analyses to a subset of students, based on several population filters, and the full set of filters are reported in Table 1. First, we focus only on students who enroll in the class during the first two weeks of its six-week run, which captures most learners who were eligible for certificates, and which excludes students playing catch-up, which may a very different experience (the results below are robust across a range of similar cut-offs). Second, we cull that sample to include only the people who finished the pre-course survey, and who were randomly assigned to see the utility value question, since these responses were going to be the focus of our prediction model. Third, we drop people who do not self-report their written English as "fluent", since we do not want the automated text analysis to simply learn differences in language skill. Finally, people who do not intend to complete the course, or who do not respond to the utility value prompt were dropped.

It is worth considering the consequences of our thorough filters. We are substantially limiting the heterogeneity in our target population, relative to the full sample of learners. This is a deliberate choice, for two reasons. First, the narrow scope allows us to model how otherwise-similar students vary in terms of their utility value, that is relatively un-confounded with other sources of variation in the broader population. Furthermore, it is worth noting that every single filter selects for students with a higher likelihood of completing the course. The resulting population are the very students who should have the highest expectations of success in the course. This makes their subsequent failure to achieve success all the more interesting.

Table 1: List of filters applied to the student population. This table shows the number of students remaining after each filter is applied, and also shows the average certification rate at each step among the remaining students.

Population Filter	Students	% Complete
Unique enrolled students	41,946	9.3%
Started in first two weeks	33,396	9.8%
Pre-Course Survey	9,862	28.9%
Saw Utility Value	4,930	28.9%
Fluent in Writing English	3,139	30.4%
Intends to Complete the Course	2,097	38.4%
Wrote more than one word	1,730	40.9%

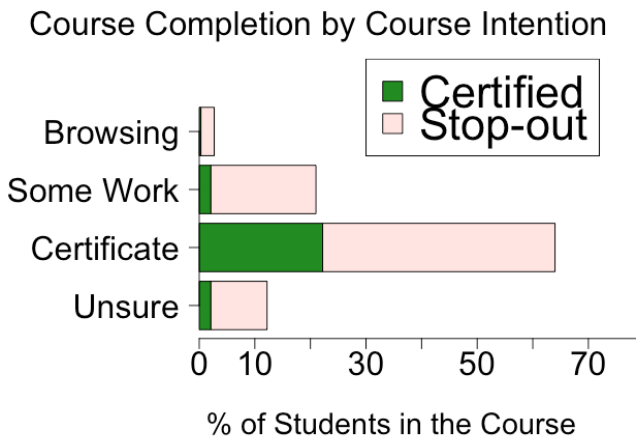


Figure 2: Course completion by student intention

The most important filter was to exclude those who did not intend to complete the course. Every HarvardX survey asks a multiple choice question about students' intentions [16]. The question asks: "People register for HarvardX courses for different reasons. Which of the following best describes you?" Students choose from four response anchors:

- Here to browse the materials, but not planning on completing any course activities (watching videos, reading text, answering problems, etc.).
- Planning on completing some course activities, but not planning on earning a certificate.
- Planning on completing enough course activities to earn a certificate.
- Have not decided whether I will complete any course activities.

Students' answers to this question are plotted in Figure 2. One clear result is that intentions matter: most of the students who certify are ones who intended to certify, and a higher percentage of students who intend to certify eventually do so, compared to those who do not intend to certify. However, the results also show that most of the students who intend to certify do not achieve that goal. The rest of this research is focused on modeling the variation in outcomes solely among those who say they intend to complete the course.

2.2 Measures

We collected two types of measures in our study. First, we collected a set of demographic and background measures from pre-course study, shown in Table 2. These measures include student age, gender, level of education, experience with MOOCs, experience with the subject material, work and student status, parental education, and residence. For the most part, students in this course were similar to students in many edX MOOCs, in that they are older than traditional college students, and very likely to have a bachelor's or advanced degree [7]. There is one notable difference from the MOOC norm: the course had a majority of female students.

In addition to collecting demographic information on participants, the primary source of data for our analyses are the open-ended responses from the utility value prompt. We combined all text in both boxes to form a single "document" for each student. We excluded anyone who wrote less than one word and, among the remainder who wrote something, the average document was 34.6 words long.

The remaining documents were cleaned and compiled using a standard NLP 9-step process. All texts were spellchecked by hand (with software assistance), all characters were converted to lowercase, all contractions were expanded, all punctuation was removed, common function words ("stopwords") were removed, every remaining word was stemmed using the standard Porter stemmer, the series of stemmed words was processed into uni- and bi-grams, a feature count matrix was constructed using all sets of features, all features which appeared in less than 2% of documents were removed.

This process produced a "feature count matrix", in which each document (i.e. each student) was assigned a row, while each n-gram feature (each word or phrase) was assigned a column, and the value of each cell represented the number of times that word or phrase was used in that document. In addition to the n-gram features, we also included two summary features for each document: word count, and the Flesh-Kincaid readability score. These data comprised the entire set of language features on which our model would train.

Table 2: Demographic characteristics of students remaining in sample. Cell contents are percentages or else response means (standard deviations in parentheses where applicable).

Demographic	Population
Gender *(% female)*	56.2%
Age	38.8 *(12.4)*
Previous MOOCs enrolled	3.2 *(3.7)*
Previous MOOCs completed	2.2 *(3.3)*
Familiar with Material (1-5)	2.8 *(1.1)*
Currently Employed	82.4%
Currently Enrolled in School	24.8%
Bachelor's Degree	77.6%
Advanced Degree	50.0%
Parent with Bachelor's Degree	57.9%
Parent with Advanced Degree	34.0%
Lives in N. America	52.3%
Lives in Europe	13.5%
Lives in Latin America	6.8%
Lives in Oceania	5.3%
Lives in Africa	5.0%
Lives in Asia	16.8%

2.3 Analytic Procedures

2.3.1 Model Estimation

Our natural language processing model was designed to predict who completed the course. The underlying statistical estimator we used was a lasso-regularized logistic regression [19]. This is a relatively simple model within machine learning that strikes a balance between bias and variance to make good out-of-sample predictions in high-dimensional environments where the solution is likely to be sparse. The lasso regularizer functions as a penalty for complexity, and the size of that penalty is determined empirically, as the penalty which minimizes prediction error in a 100-fold cross-validation loop.

2.3.2 Alternative Prediction Models

In addition to the NLP model, we tested several other prediction models using different feature sets, as benchmark comparisons. One of those benchmarks was to compare the NLP results to a model trained on demographic variables. To do that, we created a feature set from the demographics listed in Table 2, chosen based on availability and on previous research [12]. We also tested a hybrid model which used both demographic and NLP features, to see whether they complemented one another, or merely overlapped in what they said about students' likelihood of success.

We also tested the accuracy of another language-based prediction method, the LIWC, which uses a predefined dictionary to extract features from the text, rather than learning features directly from the data [15]. In our research we did not find any combination of LIWC features to be useful for out-of-sample predictions of course completion, so we do not report our LIWC analyses in detail.

3. RESULTS

The gold-standard test for model performance is prediction accuracy among "out-of-sample" cases, i.e., those cases not used in training. To accomplish this in our dataset, we used a "nested cross-validation" loop, separate from the cross-validation used to estimate the lasso parameter [20]. We used a 20-fold outer cross-validation loop, stratified to balance the proportion of successful students in each fold. We repeated the procedure five times to even out model instability, and report the averaged result as our out-of-sample prediction for each student.

Because of the imbalance in outcomes (i.e. more drop-outs than certified students) the accuracy of each set of predictions was measured as the AUC metric generated from an ROC curve [18]. The full NLP model performed well, and better than chance (AUC=56.4; $p<.001$), where students' responses to the utility value prompt predicted course completion in the hold-out sample. The demographic predictors also performed better than chance (AUC=56.1; $p<.001$), and the model put weight on only one feature – the number of previous MOOCs completed. Importantly, the signal from the demographic features and NLP features did not overlap entirely – a model trained on the combined set of features did an even better job predicting course completion than demographics alone (AUC=59.8; $p<.02$). These results confirm that the content of the students' language responses to the utility value prompt reveals an important new source of variance in their future achievement.

The selected features are listed in Table 3, along with their regularized coefficients within the model, as well as the prevalence of each feature in the documents. Each row represents a different language feature selected by the regularizer during training. The signs of the coefficients give an indication of how the model is learning to distinguish completers from non-completers, though the magnitude of the coefficients are not directly interpretable. The final column counts the prevalence of each feature in the document set for each course, as a percent of total documents. Of the 21 features in the model, 78% of students used at least one. The model made a baseline prediction for the remaining 22% of students.

Table 3: Language features selected by the lasso logistic regression model. Each row represents a different language feature selected during training. The final column counts the prevalence of each feature, as a percent of total documents.

N-Gram	Coefficient	Prevalence
get	-0.2619	5.7%
modephysic	-0.0985	2.8%
career	-0.0761	5.9%
area	-0.0761	2.4%
need	-0.0729	7.8%
practic	0.0118	5.0%
set	0.0224	4.5%
new	0.0245	18.1%
present	0.0281	2.5%
theorilearn	0.0337	6.1%
will	0.0417	76.3%
modelearn	0.0428	10.2%
see	0.0640	3.9%
profession	0.0649	5.2%
teach	0.0707	26.4%
believ	0.0735	6.1%
leadershipwill	0.1222	2.6%
impact	0.1296	3.1%
learnbetter	0.1303	2.9%
engag	0.1604	6.2%
innov	0.1816	5.0%

To give an example of how to interpret the features, the n-gram "believ" indicates all the derivatives of the word stem "believe" (i.e., "belief", "believing", "believed", etc.). In general the contents of the model suggest that students most interested in extrinsic rewards of the class (e.g. "get", "need", "career") were less likely to earn a certificate. By contrast, students who described ways in which the material might be applied on-the-job (e.g. "engage", "innovate", "impact", "teach") were most likely to follow through on their intentions and complete the course.

One interesting feature is "modephys", which indicates students using the phrase "Models of Physical Design", which is the name of a unit in the course. In fact, it is the final unit, so the negative coefficient suggests that students who were particularly interested in this unit were unlikely to wait through the rest of course to get to the desired material. This was not the case for the names of earlier units in the course (e.g.

"Theories of Learning"), though further research is needed to unpack this relationship more precisely.

4. DISCUSSION

Compared to simple benchmarks based on demographics, we find that a machine learning prediction model can learn from unstructured text to predict which students will complete an online course. We show that the model performs well out-of-sample within a single course, and better than demographic benchmarks. These results demonstrate the potential for NLP to contribute to predicting student success.

It is worth discussing some limitations to our approach. First, we are only able to make predictions for a small subset of students (i.e. those who passed all of our population filters). This excludes a lot of students who are unlikely to succeed in the course. However, those students may also be the least interested in behavioral interventions, so the applications of a prediction model among these students may be of limited value. By contrast, our work reveals insights into the minds of those who may want the most help translating their intentions into achievement.

Another clear limitation is that our research is data-intensive, and requires a large student body from which to draw language samples. However, in both MOOCs and traditional education settings, student data is increasingly a trackable entity. Schools increasingly administer student perception surveys [14] and NLP methods generate new possibilities to gain insights from student data by asking psychologically motivated open-ended questions. This research demonstrates that these questions can predict student persistence and completion over and above similar fixed response items, and this expands the scope of external validity for our results, since language might be collected from all kinds of naturally-occurring educational contexts.

REFERENCES

1. Bong, M. (2001). Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary educational psychology, 26*(4), 553-570.
2. Brooks, C., Thompson, C., & Teasley, S. (2015). A time series interaction analysis method for building predictive models of learners using log data. Paper presented at the Proceedings of the Fifth International Conference on Learning Analytics And Knowledge.
3. Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In N. Eisenberg (Ed.), *Handbook of child psychology* (Vol. 4, pp. 1017-1095). New York: John Wiley & Sons.
4. Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science, 317*(5843), 1351-1354.
5. Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of Retention and Achievement in a Massive Open Online Course. *American Educational Research Journal*, 0002831215584621.
6. Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs*, 7.
7. Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., . . . Petersen, R. (2015). HarvardX and MITx: Two Years of Open Online Courses Fall 2012-Summer 2014. doi: 10.2139/ssrn.2586847
8. Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. *(HarvardX and MITx Working Paper No. 1)*.
9. Hulleman, C. S., Durik, A. M., Schweigert, S. B., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology, 100*(2), 398.
10. Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102*(4), 880.
11. Joksimović, S., Dowell, N., Skrypnyk, O., Kovanović, V., Gašević, D., Dawson, S., & Graesser, A. C. (2015, March). How do you connect?: Analysis of social capital accumulation in connectivist MOOCs. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 64-68). ACM.
12. Kizilcec, R., & Halawa, S. (2015). Attrition and Achievement Gaps in Online Learning. *Proc. of ACM Learning at Scale, 15*, 14-15.
13. Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. Empirical Methods on Natural Language Processing (*EMNLP) 2014*, 60.
14. MET Project. (2012). Asking Student about Teaching.
15. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates, 71*, 2001.
16. Reich, J. (2014). MOOC completion and retention in the context of student intent. *EDUCAUSE Review Online*.
17. Reich, J., Tingley, D. H., Leder-Luis, J., Roberts, M. E., & Stewart, B. (2014). Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses.
18. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics, 12*(1), 77.
19. Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association, 108*(503), 755-770.
20. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 91.
21. Wen, M., Yang, D., & Rose, C. (2014). *Sentiment Analysis in MOOC Discussion Forums: What does it tell us?* Paper presented at the Educational Data Mining 2014.
22. Whitehill, J., Williams, J. J., Lopez, G., Coleman, C. A., & Reich, J. (2015). Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Stopout.
23. Yang, D., Sinha, T., Adamson, D., & Rose, C. P. (2013). *Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses.* Paper presented at the Proceedings of the 2013 NIPS Data-Driven Education Workshop.