Language to Completion: Success in an Educational Data Mining Massive Open Online Class

Scott Crossley
Georgia State U.
Atlanta, GA 30303
scrossley@gsu.edu

Danielle S. McNamara
Arizona State Univ.
Tempe, AZ, 85287
dsmcnama@asu.edu

Ryan Baker, Yuan Wang, Luc Paquette
Teachers College
Columbia University
New York, NY 10027
ryanshaunbaker@gmail.com

Tiffany Barnes
NC State Univ.
Raleigh, NC 27606
tmbarnes@ncsu.edu

Yoav Bergner
Educational Testing Service
Princeton, NJ 08541
ybergner@ets.org

ABSTRACT

Completion rates for massive open online classes (MOOCs) are notoriously low, but learner intent is an important factor. By studying students who drop out despite their intent to complete the MOOC, it may be possible to develop interventions to improve retention and learning outcomes. Previous research into predicting MOOC completion has focused on click-streams, demographics, and sentiment analysis. This study uses natural language processing (NLP) to examine if the language in the discussion forum of an educational data mining MOOC is predictive of successful class completion. The analysis is applied to a subsample of 320 students who completed at least one graded assignment and produced at least 50 words in discussion forums. The findings indicate that the language produced by students can predict with substantial accuracy (67.8 %) whether students complete the MOOC. This predictive power suggests that NLP can help us both to understand student retention in MOOCs and to develop automated signals of student success.

Keywords

Natural language processing, MOOCs, student success

1. INTRODUCTION

The sheer size of student populations in massive open online classes (MOOCs) requires educators to rethink traditional approaches to instructor intervention and the assessment of student motivation, engagement, and success [11]. As a result, a good deal of MOOC research has focused on predicting or explaining attrition and overall student success. Most research assessing student success in MOOCs has involved the examination of click-stream data. Such data provides researchers with evidence of engagement within the course and activities associated with individual course goals [6]. Additional approaches to assessing student success include the use of sentiment analysis tools to gauge students' affective states [15, 16] and individual difference measures such as student backgrounds and other demographic variables [5].

In this paper, we explore the potential for natural language processing (NLP) tools that include but also go beyond sentiment analysis to predict success in an educational data mining MOOC. Our goal is to develop an automated model of MOOC success based on NLP variables such as text length, text cohesion, syntactic complexity, lexical sophistication, and writing quality that can be used to predict class completion. Thus, in line with Koller et al. [7], we hope to better understand the language produced by MOOC students, especially differences in the language between those students that complete a course and those that do not. Using NLP variables affords the opportunity to go beyond click-stream data to examine student success and allows the personalization of predictive variables based solely on the language differences exhibited by students. Such fine-grained content analyses may allow teachers to monitor and detect evidence of student engagement, emotional states, and linguistic ability to predict success and intervene to prevent attrition.

1.1 NLP and MOOC Success

Researchers and teachers have embraced MOOCs for their potential to increase accessibility to distance and lifelong learners [7]. From a research perspective, MOOCs provide a tremendous amount of data via click-stream logs within the MOOC platform. These data can be mined to investigate student learning, student completion, and student attitudes. Typical measures include frequency of access to various learning resources, time-on-task, or attempt rates on graded assignments [14]. Less frequently mined, however, are data related to language use [15, 16].

NLP refers to the examination of texts' linguistic properties using a computational approach. NLP centers on how computers can be used to understand and manipulate natural language texts (e.g., student posts in a MOOC discussion forum) to do useful things (e.g., predict success in a MOOC). The principal aim of NLP is to gather information about human language understanding and production through the development of computer programs intended to process and understand language in a manner similar to humans [3]. Traditional NLP tools focus on a text's syntactic and lexical properties, usually by counting the length of sentences or words or using databases to compare the contents of a single text to that of a larger, more representative corpus of texts. More advanced tools provide measurements of text cohesion, the use of rhetorical devices, syntactic similarity, and more sophisticated indices of word use.

In MOOCs, the most common NLP approach to analyzing student language production has been through the use of sentiment analysis tools. Such tools examine language for positive or negative emotion words or words related to motivation, agreement, cognitive mechanisms, or engagement. For instance, Wen et al. [16] examined the sentiment of forum posts in a MOOC to examine trends in students' opinions toward the course and course tools. Using four variables related to text sentiment (words related to application, cognitive words, first person pronouns, and positive words), Wen et al. reported that students' use of words related to motivation had a lower risk of dropping out of the course. In addition, the more students used personal pronouns in forum posts, the less likely they were to drop out of the course. In a similar study, Wen et al [15] reported a significant correlation between sentiment variables and the number of students who dropped from a MOOC on a daily basis. However, Wen et al. did not report a consistent relation between students' sentiment across individual courses and dropout rates (e.g., in some courses negative words such as "challenging" or "frustrating" were a sign of engagement), indicating a need for caution in the interpretation of sentiment analysis tools.

2. METHOD

The goal of this study is to examine the potential for NLP tools to predict success in an EDM MOOC. Specifically, we examine the language used by MOOC students in discussion forums and use this language to predict student completion rates.

2.1 The MOOC: Big Data in Education

The MOOC of interest for this study is the Big Data in Education MOOC hosted on the Coursera platform as one of the inaugural courses offered by Columbia University. It was created in response to the increasing interest in the learning sciences and educational technology communities in learning to use EDM methods with fine-grained log data. The overall goal of this course was to enable students to apply each method to answer education research questions and to drive intervention and improvement in educational software and systems. The course covered roughly the same material as a graduate-level course, Core Methods in Educational Data Mining, at Teachers College Columbia University. The MOOC spanned from October 24, 2013 to December 26, 2013. The weekly course comprised lecture videos and 8 weekly assignments. Most of the videos contained in-video quizzes (that did not count toward the final grade).

All the weekly assignments were automatically graded, numeric input or multiple-choice questions. In each assignment, students were asked to conduct an analysis on a data set provided to them and answer questions about it. In order to receive a grade, students had to complete this assignment within two weeks of its release with up to three attempts for each assignment, and the best score out of the three attempts was counted. The course had a total enrollment of over 48,000, but a much smaller number actively participated; 13,314 students watched at least one video; 1,242 students watched all the videos; 1,380 students completed at least one assignment; and 710 made a post in the weekly discussion sections. Of those with posts, 426 completed at least one class assignment; 638 students completed the online course and received a certificate (meaning that some students could earn a certificate without participating in the discussion forums at all).

2.2 Student Completion Rates

We selected completion rate as our variable of success because it is one of the most common metrics used in MOOC research [17]. However, as pointed out by several researchers, learner intent is a critical issue [5, 6, 7]. Many MOOC students enroll based on curiosity, with no intention of completing the course. The increased use of entry surveys is no doubt related to this inference problem. In the present analysis, however, we do not have access to this information. Therefore, we compute completion rates based on a smaller sample of forum posters as described below. "Completion" was pre-defined as earning an overall grade average of 70% or above. The overall grade was calculated by averaging the 6 highest grades extracted out of the total of 8 assignments.

2.3 Discussion Posts

We selected discussion posts because they are one of the few instances in MOOCs that provide students with the opportunity to engage in social learning [11, 16]. Discussion forums provide students with a platform to exchange ideas, discuss lectures, ask questions about the course, and seek technical help, all of which lead to the production of language in a natural setting. Such natural language can provide researchers with a window into individual student motivation, linguistics skills, writing strategies, and affective states. This information can in turn be used to develop models to improve student learning experiences [11]. In the EDM MOOC, students and teaching staff participated in weekly forum discussions. Each week, new discussion threads were created for each week's content including both videos and assignments under sub-forums. Forum participation did not count toward student's final grades. For this study, we focused on the forum participation in the weekly course discussions.

For the 426 students who both made a forum post and completed an assignment, we aggregated each of their posts such that each post became a paragraph in a text file. We selected only those students that produced at least 50 words in their aggregated posts (n = 320). We selected a cut off of 50 words in order to have sufficient linguistic information to reliably assess the student's language using NLP tools. Of these 320 students, 132 did not successfully complete the course while the remaining 188 students completed the course.

2.4 Natural Language Processing Tools

We used several NLP tools to assess the linguistic features in the aggregated posts of sufficient length. These included the Writing Assessment Tool (WAT [9]), the Tool for the Automatic Analysis of Lexical Sophistication (TAALES [8]), and the Tool for the Automatic Assessment of Sentiment (TAAS). We provide a brief description of the indices reported by these tools below.

2.4.1 WAT

WAT was developed specifically to assess writing quality. As such, it includes a number of writing specific indices related to text structure (text length, sentence length, paragraph length), cohesion (e.g., local, global, and situational cohesion), lexical sophistication (e.g., word frequency, age of acquisition, word hypernymy, word meaningfulness), key word use, part of speech tags (adjectives, adverbs, cardinal numbers), syntactic complexity, and rhetorical features. It also reports on a number of writing quality algorithms such as introduction, body, and conclusion paragraph quality and the overall quality of an essay.

2.4.2 TAALES

TAALES incorporates about 150 indices related to basic lexical information (e.g., the number of tokens and types), lexical frequency, lexical range, psycholinguistic word information (e.g., concreteness, meaningfulness), and academic language for both single words and multi-word units (e.g., bigrams and trigrams).

2.4.3 TAAS

TAAS was developed specifically for this study. The tool incorporates a number of language-based sentiment analysis databases including the Linguistic Inquiry and Word Count database (LIWC [10]), Affective Norms for English Words (ANEW [1]), Geneva Affect Label Coder (GALC [13]), the National Research Council (NRC) Word-Emotion Association Lexicon [12], and the Senticnet database [2]. Using these databases, TAAS computes affective variables related to a number of emotions such as anger, amusement, fear, sadness, surprise, trust, pleasantness, attention, and sensitivity.

2.5 Statistical Analysis

The indices reported by WAT, TAALES, and TAAS that yielded non-normal distributions were removed. A multivariate analysis of variance (MANOVA) was conducted to examine which indices reported differences between the postings written by students who successfully completed the course and those who did not. The MANOVA was followed by stepwise discriminant function analysis (DFA) using the selected NLP indices that demonstrated significant differences between those students who completed the course and those who did not, and did not exhibit multicollinearity ($r > .90$) with other indices in the set. In the case of multicollinearity, the index demonstrating the largest effect size was retained in the analysis. The DFA was used to develop an algorithm to predict group membership through a discriminant function co-efficient. A DFA model was first developed for the entire corpus of postings. This model was then used to predict group membership of the postings using leave-one-out-cross-validation (LOOCV) in order to ensure that the model was stable across the dataset.

3. RESULTS

3.1 MANOVA

A MANOVA was conducted using the NLP indices calculated by WAT, TAALES, and TAAS as the dependent variables and the postings by students who completed the course and those who did not as the independent variables. A number of indices related to positing length, number of posts, use of numbers, writing quality, lexical sophistication, n-gram use, and cohesion demonstrated significant differences (see Table 1 for the MANOVA results). These indices were used in the subsequent DFA.

The results indicate that those who completed the course, even though course completion depended solely on success on technical assignments, tended to be better writers (i.e., received higher scores based on the essay score algorithm in WAT), to use a greater variety of words, to write more often with more words, and with greater cohesion. They also used more words relevant to the domain of the course, more concrete words, more sophisticated words, words with more associations to other words, and more common bigrams and trigrams.

3.2 Discriminant Function Analysis

A stepwise DFA using the indices selected through the MANOVA retained seven variables related to post length, lexical sophistication, the use of numbers, cohesion, and writing quality as significant predictors of whether a student received a certificate or not. These indices were *Average post lengths, Word age of acquisition, Cardinal numbers, Hypernymy standard deviation, Situational cohesion, Trigram frequency,* and *Essay score algorithm*. The remaining variables were removed as non-significant predictors.

Table 1. MANOVA Results Predicting Whether Students Completed the MOOC

Index	*F*	*η2*
Essay score algorithm	13.071**	0.039
Type token ratio	12.074**	0.037
Number of word types	11.371**	0.035
Number of posts	10.919*	0.033
Average post length	10.596*	0.032
Concreteness	10.017*	0.031
Cardinal numbers	10.081*	0.031
Trigram frequency	9.445*	0.029
Bigram frequency	8.903*	0.027
Number of sentences	8.451*	0.026
Frequency content words	8.219*	0.025
Situational cohesion	8.041*	0.025
Hypernymy standard deviation	7.643*	0.023
Word meaningfulness	7.378*	0.023
Lexical diversity	6.180*	0.019
Average word length	5.150*	0.016
Essay body quality algorithm	4.409*	0.014
Logical connectors	3.915*	0.012
Word age of acquisition	3.854*	0.012

** $p < .001$, * $p < .050$

The results demonstrate that the DFA using these seven indices correctly allocated 222 of the 320 posts in the total set, $\chi 2$ (df=1) = 46.529 $p < .001$, for an accuracy of 69.4%. For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 217 of the 320 texts for an accuracy of 67.8% (see the confusion matrix reported in Table 2 for results and F_1 scores). The Cohen's Kappa measure of agreement between the predicted and actual class label was 0.379, demonstrating fair agreement.

Table 2. Confusion matrix for DFA classifying postings

		predicted		
	actual	- Cert	+Cert	F_1 score
Whole set	- Certificate	**91**	41	0.650
	+Certificate	57	**131**	0.728
LOOCV	- Certificate	**87**	45	0.628
	+Certificate	58	**130**	0.716

4. DISCUSSION AND CONCLUSION

Previous MOOC studies have investigated completion rates though click-stream data and sentiment analysis tools. The current study adds another tool for examining successful completion of a MOOC: natural language processing. The tools assessed in this study show that language related to forum post length, lexical sophistication, situational cohesion, cardinal numbers, trigram production, and writing quality can significantly predict whether a MOOC student completed an EDM course. Such a finding has important implications for how students' individual differences (in this case, language skills) that go beyond observed behaviors (i.e., click-stream data) can be used to predict success.

Overall, the results support the basic notion that students that demonstrate more advanced linguistic skills, produce more coherent text, and produce more content specific posts are more likely to complete the EDM MOOC. For instance, students were more likely to complete the course if their posts were shorter (i.e., more efficient), used words that are less frequent or familiar (i.e., higher age of acquisition scores), used more cardinal numbers (i.e., content specific), used words that were more consistent in

terms of specificity (i.e., less variance in terms of specificity), produced posts that were more cohesive (i.e., greater overlap of ideas), used more frequent trigrams (i.e., followed expected combinations of words), and produced writing samples of higher quality (i.e., samples scored as higher quality by a automatic essay scoring algorithm). Interestingly, none of our affective variables distinguished between students who completed or did not complete the EDM MOOC. This may be the result of the specific MOOC under investigation, a weakness of the affective variables examined, or a weakness of affective variables in general.

The findings have important practical implications as well. The linguistic model developed in this paper through the DFA could be used as a prototype to monitor MOOC students and potentially identify those students who are less likely to complete the course. Such students could then be target for interventions (e.g., sending e-mails, suggesting assignments or tutoring) to improve immediate engagement in the MOOC and promote long-term completion.

The results reported in this study are both significant and extendible to similar datasets (as reported in the LOOCV results). They also open up additional research avenues. For instance, to improve detection of students who might be unlikely to complete the MOOC, follow-up models that include click-stream data could be developed and tested. Such models would likely provide additive power to detection accuracy. One concern with the current model is that it requires language samples for analysis. This suggests that NLP approaches like this one may be even more useful in classes that have activities such as collaborative chat, a feature now emerging in some MOOCs.

5. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences and National Science Foundation (IES R305A080589, IES R305G20018-02, and DRL- 1418378). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES or the NSF.

6. REFERENCES

[1] Bradley, M. M., and Lang, P. J. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. *Technical report. The Center for Research in Psychophysiology, University of Florida*.

[2] Cambria, E. and Hussain, A. 2015. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Cham, Switzerland: Springer.

[3] Crossley, S. A. 2013. Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching, 46* (2), 256-271.

[4] DeBoer, J., Stump, G. S., Seaton, D., Ho, A., Pritchard, D. E., and Breslow, L. 2013. Bringing student backgrounds online: MOOC user demographics, site usage, and online learning. *In the Proceedings of the 6th International Conference on Educational Data Mining*, 312-313.

[5] DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. 2014. Changing "Course": Reconceptualizing Educational Variables for Massive Open Online Courses. *Educational Researcher*, *March*, 74–84.

[6] Kizilcec, R. F., Piech, C., and Schneider, E. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. *In the Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 170-179.

[7] Koller, D., Ng, A., Do, C., and Chen, Z. 2013. Retention and Intention in Massive Open Online Courses. *Educause*.

[8] Kyle, K., and Crossley, S. A. in press. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly.*

[9] McNamara, D. S., Crossley, S. A., & Roscoe, R. 2013. Natural Language Processing in an Intelligent Writing Strategy Tutoring System. *Behavior Research Methods, 45* (2), 499-515.

[10] Pennebaker, J. W., Booth, R. J., and Francis, M. E. 2007. *LIWC2007: Linguistic inquiry and word count.* Austin, Texas.

[11] Ramesh, A., Goldwasser, D., Huang, B., Daume, H., and Getoor, L. 2014. Understanding MOOC Discussion Forums using Seeded LDA. *ACL Workshop on Innovative Use of NLP for Building Educational Applications,* 22-27.

[12] Saif, M., and Turney, P. 2013. Crowdsourcing a Word-Emotion Association Lexicon**,** *Computational Intelligence,* 29 (3), 436-465.

[13] Scherer, K. R. 2005. What are emotions? And how should they be measured? *Social Science Information*, 44 (4), 695-729.

[14] Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, *57*(4), 58–65.

[15] Wen, M., Yang, D. and Rose, C. P. 2014. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *In the Proceedings of the 7th International Conference on Educational Data Mining, 130-137.*

[16] Wen, M., Yang, D. and Rose, C. P. 2014. Linguistic Reflections of Student Engagement in Massive Open Online Courses. *In the Proceedings of the International Conference on Weblogs and Social Media*.

[17] Wang, Y. 2014. MOOC Leaner Motivation and Learning Pattern Discovery. *In the Proceedings of the 7th International Conference on Educational Data Mining*, 452-454.