Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion

Scott Crossley
Georgia State University
25 Park Place, Ste 1500
Atlanta, GA 30303
01+404-413-5179
scrossley@gsu.edu

Luc Paquette
University of Illinois at Urbana-Champaign
1310 S. 6th St.
Champaign, IL, 61820
01+217-300-1758
lpaq@illinois.edu

Mihai Dascalu
University Politehnica of Bucharest
313 SplaiulIndepententei
Bucharest, Romania
+40 726 111 090
mihai.dascalu@cs.pub.ro

Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+480-727-5690
dsmcnamara1@gmail.com

Ryan S. Baker
Teachers College, Columbia University
525 West 120th Street
New York, NY, 10027
01+212-678-8329
baker2@exchange.tc.columbia.edu

ABSTRACT

Completion rates for massive open online classes (MOOCs) are notoriously low. Identifying student patterns related to course completion may help to develop interventions that can improve retention and learning outcomes in MOOCs. Previous research predicting MOOC completion has focused on click-stream data, student demographics, and natural language processing (NLP) analyses. However, most of these analyses have not taken full advantage of the multiple types of data available. This study combines click-stream data and NLP approaches to examine if students' on-line activity and the language they produce in the on-line discussion forum is predictive of successful class completion. We study this analysis in the context of a subsample of 320 students who completed at least one graded assignment and produced at least 50 words in discussion forums, in a MOOC on educational data mining. The findings indicate that a mix of click-stream data and NLP indices can predict with substantial accuracy (78%) whether students complete the MOOC. This predictive power suggests that student interaction data and language data within a MOOC can help us both to understand student retention in MOOCs and to develop automated signals of student success.

Categories and Subject Descriptors

K.3.1 [**Computer Uses in Education**]: Computer-assisted Instruction (CAI); J.5 [**Computer Applications: Arts and Humanities**]: Linguistics

General Terms

Algorithms, Measurement, Performance

Keywords

MOOC, click-stream data, educational data mining, natural language processing, sentiment analysis, educational success, predictive analytics

1. INTRODUCTION

The considerable size of student populations in massive open online classes (MOOCs) requires educators and administrators to rethink traditional approaches to instructor intervention and the manner in which student engagement, motivation, and success is assessed [34]. As a result of differences between traditional classes and MOOCs, a new research agenda focusing on predicting or explaining attrition and overall student success in MOOCs has emerged. The most common data available for such analyses is click-stream data (i.e., student interactions within the MOOC software). Such data provides researchers with evidence of engagement within the course and activities associated with individual course goals [22] and, because of this, most research assessing student success in MOOCs has involved the examination of click-stream data. Additional approaches used in recent studies to assessing student success include the use of sentiment analysis tools to gauge students' affective states [42, 43], linguistic features that measure the sophistication and organization of student writing within a MOOC [9], and individual difference measures such as student background and other demographic variables [16].

In this paper, we combine NLP and click-stream data approaches to examine success in an educational data mining MOOC as called for by Crossley and colleagues [9]. Thus, unlike prior research, we investigate the potential for NLP indices related to text length, social collaboration, sentiment analysis, text cohesion, syntactic complexity, lexical sophistication, and writing quality in conjunction with click-stream data such as lecture viewing and page views to predict MOOC completion. We hypothesize that developing complete and predictive understandings of student outcomes requires multiple sources of data and a variety of approaches because learning is not simple, but rather a complex process with multiple layers and time-scales. Relying on a single data sources, be it language information or click-stream data, limits the potential for understanding student differences, especially when automated streams of data from different learner dimension are readily available.

As such, this paper's aim is to create an automated model of MOOC success based on both click-stream and NLP indices. We focus on student completion rate because it is an important component of student success [20], not only in itself but because it predicts participation in a scientific community of practice after taking the MOOC [45]. The NLP tools we use in conjunction with traditional click-stream data open a wide array of opportunities for better understanding student success. Our objective is to examine the efficacy of combined approach in predicting the probability of MOOC course performance and completion. The long-term goal of this research is to inform interventions that provide personalized feedback in terms of language use and system interaction to MOOC students or their teachers in order to increase completion rate, as well as to increase scientific understanding of the factors associated with MOOC completion.

1.1 Click Stream Data in MOOCS

Researchers and instructors have embraced MOOCs for their potential to increase accessibility to distance and lifelong learners [24]. From a research perspective, MOOCs provide a tremendous amount of data via click-stream logs which provide detailed records of the students' interactions with course content, including video lectures, discussion forums, assignments, and additional course content, within the MOOC platform. These data can be mined to investigate student completion [1, 2, 18, 19, 23, 39, 46] and student engagement [22, 38].

Typical measures computed from click-stream data and used in MOOC analyses includes variables related to the timing of actions, counts of the different possible types of actions, forum interactions and assignments attempts [37]. Timestamps in the click-stream have been used to compute average delays in watching lectures as they become available [38], time difference between assignment submissions and assignment deadlines [2. 38] or to identify students who are lagging behind [18]. Student interactions with course forums can be used to determine the number of times each student read different forum thread or contributed by posting on the forums [1]. Click-stream data has also been used to examine how students interact with course assignments by calculating the number of distinct problems a student attempts and the average number of submissions by problem [2]. Data about interaction with video lectures can be used to compute the percentage of available lectures that the student has watched [1, 19].

Click-stream has been used to build models predictive of whether a student will drop out of the course in the following weeks or fail to obtain a certificate of completion of the course [46]. Researchers have applied machine learning techniques, including logistic regression [2, 46] and hidden Markov models [1], in order to create models able to detect at-risk students early in the course. In addition, researchers have been able to develop models of student success based on click-stream data that can generalize to new courses [2, 46]. For instance, Balakrishnan [1] used click-stream data to examine the probability that a student would stay in the course for the following week. They reported high predictive power for students that remained in the class (F1 scores of .888), but low ability to identify the students that dropped (F1 scores of .115). A similar approach has also been applied to study student retention in e-Learning courses using learning management systems (LMS). Accuracy rates in these studies reach around 90-95% for data that includes course assignments performance (i.e., quiz grades and assignment completion) along with demographic information such as gender and work experience and click-stream data [26, 27].

In addition to the creation of predictive models of dropout, click-stream data has been used to study student engagement in MOOCs. Sharma and colleagues [38] studied the behavior of students in relation to pre-defined engagement profiles. In their study, they used data about graded assignment submissions to classify students as either an *active student* or a *viewer*. They further subdivided active students into categories indicating whether a student earned a certificate with distinction, completed the course, or failed and subdivided viewers into active viewers, passive viewers, wiki-users, dropout and completer. Kizilcec and colleagues [22] applied clustering algorithms to click-stream data in order to identify four engagement profiles, completing, auditing, disengaging and sampling, based on students' weekly behavior in the course. As such, click-stream data has been shown to be a useful tool for mining and modeling student behavior in MOOCs in that it can predict completion, continued participation, and engagement.

1.2 NLP and MOOC Analysis

Less frequently mined than click-stream data (at least in the context of MOOCs) are data related to language use using NLP tools [7, 42, 43]. NLP refers to the computational examination of texts' linguistic properties. 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) [8]. 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 structures, syntactic similarity, social collaboration, sentiment analysis, topic development, and sophisticated indices of word usage.

A number of analyses have used NLP techniques to investigate elements of MOOCs unrelated to student success. For instance, Elousazizi [17] used linguistic markers of point of view to determine that MOOC participants showed low levels of cognitive presence and engagement. Moon, Potdar, and Martin [32] used emotion terms and semantic similarity among participants to identify student leaders. Lastly, Chaturvedi, Goldwasser, and Daume [3] used words related to assessment, technical problems, politeness, and requests to predict instructor intervention.

The most common NLP approach to analyzing student language production in MOOCS 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. Of relevance to the current study are sentiment analyses that predict student success such as Wen et al. [43], who examined the sentiment of forum posts in a MOOC to examine trends in students' opinions toward the course and overall success. Wen et al. reported that students who used 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 [42] 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.

More recently, Crossley et al. [9] used a number of NLP tools to examine the sophistication of language in forum posts and whether or not the language used was predictive of MOOC completion. Crossley et al. drew from a number of linguistic variables related to text cohesion, syntactic complexity, and lexical sophistication. Their analysis indicated that language related to forum post length, lexical sophistication, situational cohesion, cardinal numbers, trigram production, and writing quality were significantly predictive of whether a MOOC student completed the course or not (with an accuracy of 67%). Their findings supported the notion that students who demonstrate more advanced linguistic skills, produce more coherent text, and produce more content specific posts are more likely to complete a MOOC.

METHODS

The goal of this study is to examine the potential for NLP tools in conjunction with click-stream data to predict success in a MOOC. Specifically, we examine student interaction within the system and the language used by MOOC students in discussion forums to predict student completion rates. We do not investigate performance variables such as quiz grades and homework scores because completion and performance data strongly overlap.

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 in 2013 as the inaugural MOOC offered by Teachers College Columbia University. It is the same MOOC investigated by Crossley and colleagues [9]. The MOOC 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 a range of EDM methods 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 ran 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, and composed of 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. The best score out of the three attempts was counted towards the final grade. The course had a total enrollment of over 48,000 in the time before the official course end, 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 discussion forums. 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 earned a certificate without participating in the discussion forums at all).

1.3 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 [19]. 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.

1.4 Discussion Posts

Discussion posts are of interest within research on student participation in MOOCs because they are one of the few instances in x-MOOCs that provide students with the opportunity to engage in social learning [34, 43]. Discussion forums provide students with a platform to exchange ideas, discuss the 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 [34]. 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 who 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 students’ language using NLP tools. Of these 320 students, 132 did not successfully complete the course while the remaining 188 students completed the course.

1.5 Click-Stream Data

1.5.1 Lectures

Click-stream data allowed us to create variables related to the students' interactions with video lectures during the course. First, we computed a global variable that indicated the average number of times, per active week (i.e., only the weeks in which the student was actively involved in the MOOC), where the student interacted with the lectures by accessing (viewing or downloading) a lecture ($M = 12.51$; $SD = 6.59$). Second, we computed the percentage of available lectures that a student had accessed at the end of each week (e.g., the percentage of 1st week lectures the student had accessed by the end of week 1, the percentage of 1st and 2nd week lectures that were accessed by the end of week 2). We refer to this variable as the student's *weekly lecture coverage*. This coverage was averaged across each week of student activity. Three versions of this variable were created depending on whether the student viewed the videos online ($M = 0.66$; $SD = 0.28$), downloaded the videos ($M = 0.58$; $SD = 0.35$), or accessed the videos using either method ($M = 0.81$; $SD = 0.20$).

1.5.2 Forum Interaction

We also used the click-stream data to create variables related to the students' interactions with discussion forums. From this data, we computed averages, per week, of how much the student was active by measuring how often a student accessed a forum ($M = 21.91$; $SD = 18.13$), created a post ($M = 2.11$; $SD = 1.46$), or commented on an existing post ($M = 1.91$; $SD = 1.15$). In addition, we looked at students' interactions with the forum reputation system. When using the forum, students had the option of providing upvotes and downvotes on posts and comments to indicate whether it was useful or not. We computed the average number of times, per week of activity, that a student upvoted

(*M* = 3.01; *SD* = 3.36) or downvoted (*M* = 1.54; *SD* = 1.12) a post or a comment on the forums.

1.5.4 Page Views

We computed two variables related to page views in the course. First, we computed a variable indicating the average number of times, per active week, that a student accessed any of the course's web pages (*M* = 81.53; *SD* = 45.54). Second, we computed a similar variable indicating the average number of times, per active week, that a student accessed the course's syllabus (*M* = 2.45; *SD* = 1.65).

1.5.5 Assignments

Finally, three variables were computed relative to the students' interactions with the course's assignments. First, we computed the average number of times, per active week, that a student submitted an assignment (*M* = 4.38; *SD* = 2.22). We also looked at how quickly students completed the assignments. We computed variables indicating the average time (in seconds), per active week, before an assignment's submission deadline that a student submitted first attempts (*M* = 645,801.7; *SD* = 356,234.4) and last attempts (*M* = 608,993.1; *SD* = 352,286.8).

1.6 Natural Language Processing Tools

We used several NLP tools to assess the linguistic features in the aggregated posts that were of a sufficient length (50 words). These included the Writing Assessment Tool (WAT [29]), the Tool for the Automatic Analysis of Lexical Sophistication (TAALES [25]), the Tool for the Automatic Analysis of COhesion (TAACO [10]), ReaderBench (RB [11]), and The SEntiment ANalysis and Cognition Engine (SEANCE). The latter three tools were not used in Crossley et al. [9]. We provide a brief description of the indices reported by these tools below.

1.6.1 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.

1.6.2 TAALES

TAALES is a freely available NLP tool that reports on a number of indices related to lexical sophistication. 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).

1.6.3 TAACO

TAACO is a freely available NLP tool that reports on a number of indices related to text cohesion. TAACO incorporates over 150 classic and recently developed indices related to text cohesion. The cohesion indices reported by TAACO evenly focus on local cohesion, global cohesion, and overall text cohesion. Local cohesion refers to cohesion at the sentence level (i.e., cohesion between smaller chunks of text such as noun overlap between sentences or linking sentences through connectives) while global cohesion refers to cohesion between larger chunks of text such as paragraphs (e.g., noun overlap between paragraphs in a text). Overall text cohesion refers to the incidence of cohesion features in an entire text, but not in comparison to other parts of the text (e.g., lexical diversity which is calculated as the repetition of words across a text).

1.6.4 ReaderBench

ReaderBench (RB) is an automated software framework that integrates text mining techniques, advanced natural language processing, and social network analysis tools in order to predict and assess learner comprehension [11, 12]. *ReaderBench* is based on a cohesion-based representation of the discourse that can be applied to language from different educational sources (e.g., Computer Supported Collaborative Learning – CSCL – conversations conducted in MOOC, textual materials, or metacognitive explanations) [14]. In terms of CSCL analyses, RB introduces two computational models used to automatically assess collaboration based on social knowledge building and voice inter-animation [15]. In addition, *RB* provides the bases for a multi-dimensional analysis of textual complexity applied on learner contributions that covers a multitude of factors ranging from classic readability formulas, surface metrics derived from automatic essay grading techniques, morphology, syntax indices, as well as a particular emphasis on semantics [13].

1.6.5 SEANCE

The SEntiment ANalysis and Cognition Engine (SEANCE) is a freely available sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. Unlike other sentiment analysis tools commonly used data mining studies (i.e., LIWC [33]), SEANCE is freely available and contains part of speech (POS) tags and valence features. SÉANCE, TAALES, and TAACO are available at http://www.kristopherkyle.com/seance.html. SEANCE indices are taken from freely available source databases such as SenticNet [5, 6] and EmoLex [30, 31]. For many of these dictionaries, SEANCE also provides a negation feature (i.e., a contextual valence shifter) that ignores positive terms that are negated. The negation feature, which is based on [21], checks for negation words in the three words preceding a target word. SEANCE also includes the Stanford part of speech (POS) tagger [41] included in Stanford CoreNLP (Manning et al., 2014). The POS tagger allows for POS tagged specific indices for nouns, verbs, and adjectives. POS tagging is an important component of sentiment analysis because unique aspects of sentiment may reside more strongly in adjectives or in verbs and adverbs [20, 40]. The SEANCE tool can report on almost 3,000 indices, but such a large number of indices can be unwieldy. Thus, SEANCE also reports on 20 component scores derived from the SEANCE indices.

1.7 Analyses

The click-stream variables and the indices reported by WAT, TAALES, TAACO, Reader Bench and SEANCE that yielded non-normal distributions were removed because they violated statistical assumptions. A multivariate analysis of variance (MANOVA) was conducted to examine which variables 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.

Table 1: MANOVA Results Predicting Whether Students Completed the MOOC

Index	*F*	$\eta2$
Average weekly lecture coverage (online views)	64.049**	0.169
Time before deadline for first attempt	20.920**	0.062
Time before deadline for last attempt	18.879**	0.056
Average page views	16.252**	0.049
Average number of lectures accessed	15.173**	0.046
High school essay score	12.800**	0.039
Certainty component score	12.248**	0.037
Type token ratio	11.962**	0.036
Coverage of top 10 topics	11.601**	0.035
New threads started	11.098**	0.034
Number of words produced	11.137**	0.034
Number of contributions	10.835**	0.033
Average post length	10.102*	0.031
Cardinal numbers used	9.789*	0.030
Concreteness	9.892*	0.030
Overall social knowledge building	9.284*	0.029
Tri-gram frequency	9.309*	0.029
Degree of voice inter-animation	9.222*	0.028
Bi-gram frequency	8.869*	0.027
Celex content word frequency (written) SD	7.774*	0.024
Incidence of periods	7.642*	0.024
Situational cohesion	7.732*	0.024
Word meaningfulness	7.549*	0.023
Word hypernymy SD	7.387*	0.023
Semantic similarity between paragraphs	7.397*	0.023
CELEX content word frequency (spoken)	6.294*	0.020
Lexical diversity M	6.373*	0.020
Unique named entities per paragraph	6.002*	0.019
Number of upvotes	5.973*	0.019
Number of entities per sentence	5.043*	0.016
Average number of order per paragraph	4.985*	0.016
Verb polarity	4.566*	0.014
Average forum reads	4.066*	0.013
Lexical diversity MTLD	3.996*	0.012

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

2. RESULTS

2.1 MANOVA

A MANOVA was conducted using the click-stream and the NLP indices calculated by WAT, TAALES, TAACO, RB, and SEANCE as the dependent variables and the postings by students who completed the course and those who did not as the independent variables. The strongest predictors were click-stream variables followed by NLP variables. A number of click-stream variables related to videos viewed, time before turning in assignments, page views, and lectures views showed strong effect sizes. The NLP indices that showed strong effect sizes came from a range of different tools and measured a variety of language constructs including writing proficiency, fluency, local and global cohesion, sentiment, use of numbers, lexical sophistication, n-gram use, named entities, and social collaboration (see Table 1 for the MANOVA results). These indices were used in the subsequent DFA.

2.2 Discriminant Function Analysis

A stepwise DFA using the indices selected through the MANOVA retained ten variables related to videos viewed, number of entities in the posts, post length, lexical sophistication, writing proficiency, time before deadline for completing assignments, cohesion between paragraphs, and certainty (see Table 2 for indices and standardized co-efficients). The remaining variables were removed as non-significant predictors.

Table 2: Discriminant Function Co-efficients

Index	Co-efficient
Average weekly lecture coverage (online views)	0.688
Number of entities per sentence	-0.358
Average post length (sentences)	-0.329
Tri-gram frequency	0.324
High school essay score	0.275
Word hypernymy SD	-0.257
Time before deadline for first attempt	0.229
Semantic similarity between paragraphs	0.222
Certainty component score	0.201
Word meaningfulness	-0.194

Table 3: Confusion matrix for DFA classifying postings

		Predicted		
	Actual	- Cert	+Cert	F_1 score
Whole set	- Certificate	**99**	32	0.773
	+Certificate	39	**148**	0.766
LOOCV	- Certificate	**95**	36	0.754
	+Certificate	40	**147**	0.750

The results demonstrate that the DFA using these 10 indices correctly allocated 247 of the 320 participants in the total set, $\chi2$

(df = 1) = 93.893 $p < .001$, for an accuracy of 77.7%. For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 242 of the 320 participants for an accuracy of 76.1% (see the confusion matrix reported in Table 3 for results and F_1 scores). The Cohen’s Kappa measure of agreement between the predicted and actual completion rate was 0.543 when LOOCV was conducted, demonstrating moderate agreement.

3. DISCUSSION

Previous MOOC studies have investigated completion rates though click-stream data or NLP techniques. The current study combines these two techniques to examine successful MOOC completion, choosing these variables, instead of performance variables and demographic information, in order to develop a model that is actionable and non-tautological. The findings indicate that variables based on click-stream data were the strongest predictors of MOOC completion but that NLP variables were also predictive. The click-stream variables that were most predictive included the weekly lecture coverage (online views) and how early students submitted their assignments. In terms of language features, indices related to the number of entities in a forum post, the post length (average number of sentences), the overall quality of the written post, lexical sophistication, cohesion between posts, and word certainty were also strong predictors of MOOC completion. Such findings have important implications for how students’ interactions within the MOOC (i.e., observed behaviors) and individual differences (in this case, language skills) can be used to predict success.

The results indicate that those who completed the course interacted more within the system on average (not cumulative) and were more active in forums. For instance, in terms of course interaction, students who had a greater weekly lecture coverage (online views), turned in their assignments earlier, viewed more pages, and accessed lectures more often were more likely to complete the course. In addition, in terms of forum participation, students who created more forum threads and contributed more posts, gave upvotes more often, and read more forum posts were more likely to complete the MOOC. These findings hold even though completion depended solely on success on technical assignments.

The linguistic results demonstrate that students who are more likely to complete the MOOC produced higher quality writing samples (as indicated by essay scoring algorithms) and, in some cases, more sophisticated language as evidenced by less meaningful (i.e., words with fewer associations), less concrete words, and longer sentences (i.e., fewer sentences per post). However, students who were more likely to complete the MOOC also used more frequent words, bigrams, and trigrams. For a cohesion perspective, students who were more likely to complete the MOOC produced writing samples that were more coherent with lower type-token ratios (i.e., greater repetition of words), more cohesion between paragraphs (i.e., semantic similarity between paragraphs), and greater use of connectives (i.e., order connectives). From a content perspective, students more likely to complete the MOOC used a greater incidence of cardinal numbers and fewer named entities indicating a focus on numbers and not names. Successful students were also more likely to stay on topic (i.e., reporting higher on-topic relatedness and content coverage scores) and were more likely to exhibit social collaboration (as seen in cohesion building graphs). Lastly, from a sentiment perspective, successful students expressed more certainty and used more negative verbs (e.g., *neglect, dislike*).

To illustrate these findings, we present excerpts from two students in the MOOC (students 97211 and 1780650, see Table 4 for excerpts). Student 97211 completed the certificate (an average score of 1), while student 1780650 did not complete the course or receive a certificate (average score of .50). The linguistic and click-stream data for each participant trend in a manner similar to the co-efficients reported in the DFA (see Table 2 for the DFA co-efficients). Mean scores for each of the linguistic indices and click-stream variables are reported in Table 5).

Table 4: Text excerpts from students that completed and did not complete the EDM MOOC

Student 97211 (completed the MOOC)
Yes, from the goal of finding temporal patterns (i.e. sequences), sequences of length 1 don't make much sense. After all, those were not included in the final slides of the GSP-algoritm part of the lecture. On the other hand: Why starts the algorithm with calculating support for the length-1-sequences? The algorithm ought to start with length 2 then. To understand the algorithm, I just took the letters as symbols, while you interpreted 'a' as "GAMING and BORED 5:05:20", 'd' as "GAMING and BORED 5.06:20" and so on. But what then is the meaning of triplet 'aad'? Two times gaming+bored at the same time and once again a minute later??
Student 1780650 (did not complete the MOOC)
Hi, In the video lecture W5V3, the rule: If the student took advanced data mining, then the student took intro to statistics. The Support/Coverage = 2/11 (why not 7/11) Because the implication A=>B if B is true than the implication is always true (The student took advanced data mining doesn't matter anymore) And from the definition of Support/Coverage: Number of data points that fit the rule? (Fit means that the true is true? right?) Any ideas?

Table 5: Mean scores for selected students

Index/Variable	Student 97211	Student 1780650
Average videos viewed	1	0.187
Number of entities per sentence	0.848	0.591
Average post length	1.344	2.667
Tri-gram frequency	1.822	1.806
High school essay score	6	4
Word hypernymy SD	2.862	2.866
Time before deadline for first attempt	954201.25	-112404.5
Semantic similarity between paragraphs	0.312	0.243
Certainty component score	0.191	0.19
Word meaningfulness	96.004	100.691

The excerpts provide some indication of the linguistic differences between students who completed the MOOC and those that did not. For instance, student 97211, who completed the MOOC, used more certainty words (e.g., *yes, ought to*) and used longer sentences (i.e., fewer sentences per post). Student 1780650, who did not complete the MOOC, used short sentences, more meaningful words (e.g., *student, true, number, idea*), and less frequent trigrams (e.g., *true is true*). Such differences likely informed the writing quality algorithm as well, which indicated that student 97211's posts were scored a 6 while student 1780650's posts were scored a 4. However, it should be noted that these excerpts do not represent the entire population or even the entire sample size from each participant. Rather, the excerpts provide an illustration of linguistic trends that are likely captured on a much larger scale by the machine learning algorithms we employed in our analyses.

In terms of click-stream variables, student 97211, who completed the MOOC and received a certificate, viewed all the videos available on the MOOC, while student 1780650, who did not complete the MOOC, only viewed 19% of the videos. In addition, student 97211 had a positive time before the deadline for first attempt while student 1780650 had a negative time before the deadline for first attempt. These findings demonstrate that more successful students viewed more videos and turned in their assignments early.

Including both linguistic indices and click-stream variables offers an improvement over models based solely on linguistic features of about 10%, although we did use a greater number of linguistic features than previous research [9]. This provides some support for our hypothesis that better understanding student outcomes requires multiple sources of data and a variety of approaches because learning is not simple, but rather a complex process with multiple layers and time-scales. Relying on a single data source, such as language features alone, limits the potential for understanding student differences. However, combine language features with click-stream data improves both our understanding of student success and our predictive models

The findings from this study have practical implications. As noted in previous studies [9], the models developed in this paper could be used to monitor MOOC students and potentially identify those students who are less likely to complete the course. Such students could then be targeted for interventions (e.g., sending e-mails, suggesting assignments or tutoring) to improve immediate engagement in the MOOC and promote long-term completion. Also as noted previously [9], models such as those developed here require students to produce language samples in order to conduct the NLP analysis. Language samples are not always readily available and not always required in MOOCs indicating that NLP is not a necessary requirement for automatically modeling student success. However, knowing that models need to be dynamic and knowing that the number and types of interactions available in a MOOC are not infinite, NLP techniques would seem a strong counterpart to click-stream data for developing rigorous models that are domain independent. MOOCs that want to take fuller advantage of models such as those reported here should offer and/or require students opportunities to interact in forums or produce language samples beyond forum spaces. Such samples could include collaborative chats or written assignments that can be automatically scored such as lecture summarizations. Once MOOCs start to involve a greater amount of language production, the models discussed here can be tested across class domains and the generalizability of combined models can be assessed. We presume that since tacit features of language (like many of those examined in this study) will remain stable across learners regardless of topic, models informed through both NLP and click-stream data can be expected to prove to be reliable across domains, although this remains to be tested in domains beyond the one studied in this paper (Educational Data Mining). Thus, future iterations of this work will examine MOOCs in other domains, MOOCs taught by other instructors, and MOOCs using other platforms.

4. 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.

5. REFERENCES

[1] Balakrishnan, G., & Coetzee, D. 2013. Predicting Student Retention in Massive Open Online Courses Using Hidden Markov Models. *Electrical Engineering and Computer Sciences University of California at Berkeley*.

[2] Boyer, S., & Veeramachaneni, K. 2015. Transfer Learning for Predictive Models in Massive Open Online Courses. *In the Proceedings of the International Conference on Artificial Intelligence in Education*.

[3] 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*.

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

[5] Cambria, E., Havasi, C., & Hussain, A. 2012. SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In G. M. Youngblood & P. M. Mcarthy (Eds.), *Proceedings of the 25th Florida artificial intelligence research society conference* (pp. 202-207).

[6] Cambria, E., Speer, R., Havasi, C., & Hussain, A. 2010. SenticNet: A publicly available semantic resource for opinion mining. In C. Havasi, D. Lenat, & B. Van Durme (Eds.), *Commonsense Knowledge: Papers from the AAAI Fall Symposium* (pp. 14-18).

[7] Chaturvedi, S., Goldwasser, D., & Daume, H. 2014. Predicting instructor's intervention in MOOC forums. *Proceedings of the 52nd Meeting of the Association for Computational Linguistics*.

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

[9] Crossley, S. A., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. 2015. Language to completion: Success in an educational data mining massive open online class. In Santos, O. C., Boticario, J. G., Romero, C., Pechenizkiy, M., Merceron, A., Mitros, P., Luna, J. M., Mihaescu, C., Moreno, P., Hershkovitz, A., Ventura, S., & Desmarais, M. (eds.) *Proceedings of the 8th International Conference on Educational Data Mining.* (pp. 388-392).

[10] Crossley, S. A., Kyle, K., & McNamara, D. S. in press. The Tool for the Automatic Analysis of Text Cohesion

(TAACO): Automatic Assessment of Local, Global, and Text Cohesion. *Behavior Research Methods.*

[11] Dascalu, M., 2014. Analyzing discourse and text complexity for learning and collaborating*, Studies in Computational Intelligence*. Springer, Switzerland.

[12] Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A., 2014. Mining texts, learners productions and strategies with ReaderBench. In *Educational Data Mining: Applications and Trends*, A. Peña-Ayala Ed. Springer, Switzerland, 335–377.

[13] Dascalu, M., Stavarache, L.L., Trausan-Matu, S., Dessus, P., & Bianco, M., 2014. Reflecting Comprehension through French Textual Complexity Factors. In *26th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2014)* IEEE, Limassol, Cyprus, 615–619.

[14] Dascalu, M., Trausan-Matu, S., Dessus, P., & Mcnamara, D.S., 2015. Discourse cohesion: A signature of collaboration. In *5th Int. Learning Analytics & Knowledge Conf. (LAK'15)* ACM, Poughkeepsie, NY, 350–354.

[15] Dascalu, M., Trausan-Matu, S., Mcnamara, D.S., & Dessus, P., in press. ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning*.

[16] 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.

[17] Elouazizi, N. 2014. Point of view mining and cognitive presence in MOOCs: A (computational) linguistic perspective. *Proceedings of the Empirical Methods in Natural Language Processing Workshop*, 32-37.

[18] Halawa, S., Greene, D., & Mitchell, J. 2014. Dropout Prediction in MOOCs Using Learner Activity Features. *Experiences and Best Practices in and Around MOOCs*, 7.

[19] He, J., Bailey, J., Rubinstein, B.I., Zhang, R. 2015. Identifying At-Risk Students in Massive Open Online Courses. *In Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[20] Hu, M., & Liu, B. 2004. Mining and summarizing customer reviews. In W. Kim & R. Kohavi (Eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168-177).

[21] Hutto, C. J., & Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar & P. Resnick (Eds.), *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 216-225).

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

[23] Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. 2014. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. *The 2014 Conference on Empirical Methods on Natural Language Processing*.

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

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

[26] Lauria, E.J., Baron, J.D., Devireddy, M., Sundararaju, V., & Jayaprakash, S.M. 2012. Mining Academic Data to Improve College Student Retention: An Open Source Perspective. *In Proceedings of the 2nd Conference on Learning Analytics and Knowledge*, 139-142.

[27] Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. 2009. Dropout Prediction in e-Learning Courses Through the Combination of Machine Learning Techniques. *Computers & Education*, 53(3), 950-965.

[28] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60).

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

[30] Mohammad, S. M., & Turney, P. D. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26-34). Association for Computational Linguistics.

[31] Mohammad, S. M., & Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436-465.

[32] Moon, S., Potdar, S., & Martin, L. 2014. Identifying student leaders from MOOC discussion forums through language influence. *Proceedings of the Empirical Methods in Natural Language Processing Workshop*, 15-20.

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

[34] 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.

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

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

[37] 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.

[38] Sharma, K., Jermann, P., & Dillenbourg, P. 2015. Identifying Styles and Paths Toward Success in MOOCs. *In the Proceedings of the 8th International Conference on Educational Data Mining*, 408-411.

[39] Taylor, C., Veeramachaneni, K., & O'Reilly, U.M. 2014. Likely to Stop? Predicting Stopout in Massive Open Online Courses. *arXiv preprint*, arXiv:1408.3382.

[40] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

[41] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.

[42] 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*.

[43] 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*.

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

[45] Wang, Y.E., Paquette, L., Baker, R. 2015. A Longitudinal Study on Learner Career Advancement in MOOCs. *Journal of Learning Analytics, 1* (3), 203-206.

[46] Whitehill, J., Williams, J.J., Lopez, G., Coleman, C.A., & Reich, J. 2015. Beyond Predictions: First Steps Toward Automatic Intervention in MOOC Student Dtopout. Available at SSRN 2611750.