121st ASEE Annual Conference & Exposition
Indianapolis, IN
June 15-18, 2014
Paper ID #10058

Mining Student-Generated Textual Data In MOOCS And Quantifying Their Effects on Student Performance and Learning Outcomes

Dr. Conrad Tucker, Pennsylvania State University, University Park
Barton K. Pursel, The Pennsylvania State University

Barton K. Pursel, Ph.D., is a Research Project Manager at the Pennsylvania State University, focusing on the intersection of technology and pedagogy. Barton works collaboratively with faculty across disciplines to explore how emerging technologies and trends, such as MOOCs, digital badges, and learning analytics, impacts both students and instructors.

Anna Divinsky

Mining Student-Generated Textual Data in MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes

Abstract

Massive Open Online Courses (MOOCs) are freely available courses offered online for distance based learners who have access to the internet. The tremendous success of MOOCs can in part, be attributed to their global availability, enabling anyone in the world to sign up/drop courses at any time during the course offerings. A single course enrollment in MOOCs can range between 10,000 to 200,000 students, hereby providing a potentially rich venue for large scale digital data (e.g., student course comments, temporal and geo-location data, etc.). However, despite the overabundance of digital data generated through MOOCs, research into how student interactions in MOOCs translate to student performance and learning outcomes is limited.

The objective of this research is to mine student-generated textual data (e.g., online discussion forums) existing in MOOCs in order to quantify their impact on student performance and learning outcomes. Student performance is quantified based on grades attained in course homework assignments, quizzes and examinations. Similar to in-class learning environments, students enrolled in MOOCs often self-organize and form learning groups, where course topics and assignments can be discussed. One of the major benefits of MOOC data is that student networks and discussion therein are digitally stored and readily available for data mining/statistical analysis. The proposed methodology employs robust natural language processing techniques and data mining algorithms to quantify temporal changes in student sentiments relating to course topics and instructor clarity. Researchers aim to determine whether textual content (e.g., quality VS quantity of student forum discussions) expressed through MOOCs can serve as leading indicators of student performance in MOOCs. A case study involving the *Introduction to Art: Concepts and Techniques* offered by Penn State University through the Coursera platform, is used to validate the proposed methodology.

1. Introduction

The advent of low cost computing and network infrastructure has increased the accessibility and affordability of the internet. In the United States alone, internet usage has increased from 43% to over 81% (2012), well above the average of 73% (2012) for other developed nations [1]. Developing nations are also seeing a surge in internet accessibility and usage, although the disparity is greater, ranging from 1%-50%, depending on the country [1]. The result has been a paradigm shift in the accessibility of educational resources. Online education environments continue to gain popularity, both at traditional brick and mortar University establishments (e.g., MIT's OpenCourseWare [2]) and emerging virtual education environments that conduct a majority of their educational objectives online (e.g., University of Phoenix[3]). Massive Open Online Courses (MOOCs) are freely available courses offered online, enabling anyone in the world with

access to internet, to sign up/drop courses at any time during the course offerings[4]. The global reach and quality of content found in MOOCs has resulted in course offerings that typically comprise of anywhere between 10,000 to over 200,000 students per course offering [5]. Online course platforms such as Cousera have over 5 million students registered on their site. Students in a typical MOOC have the ability to access course assessments (e.g., assignments, quizzes), discuss course learning objectives and outcomes with other students (e.g., through an online forum) or directly communicate with an instructor (e.g., through private email messages or through public MOOC forums). As a result of multiple avenues of communication, the amount of data (primarily textual in nature) generated by students is substantial, hereby providing a potentially rich venue for large scale digital data (e.g., student course comments, temporal and geo-location data, etc.). However, despite the overabundance of textual digital data generated through MOOCs, research into how student interactions in MOOCs translates to student performance and learning outcomes has been limited. Unlike traditional brick and mortal university establishments, online education systems such as MOOCs provide the unique opportunity for instructors and researchers to capture and model student feedback, engagement and interest in course topics, as the course progresses (e.g., through digital interactions with other students, instructors, etc.). While brick and mortal university courses typically have a component of the course that enables students to communicate digitally, students in online only courses such as MOOCs are constrained by their geographical location, making virtual interaction the primary mode of communication.

The objective of this research is to mine student-generated textual data (e.g., online discussion forums) existing in MOOCs in order to investigate the relationship between student sentiment (expressed textually in MOOCs and quantified using advanced data mining/natural language processing algorithms) and student performance in the course (quantified based on grades attained in course homework assignments, quizzes and examinations).

2. Literature Review

2.1 Massively Open Online Courses

The Massive Open Online Course, or MOOC, is a relatively new development in education. The roots of MOOCs can be traced back to similar initiatives such as MIT's *Open Courseware Initiative*, the Open University's *OpenLearn*, and other open educational resource (OER) efforts throughout the early 2000s. The ideals of many of these OER efforts, as well as MOOCs, is to provide free access to knowledge for everyone, regardless of geographic, demographic or economic constraints [6]. *Connectivism and Connective Knowledge*, a course by George Siemens and Stephen Downes, was arguably the first MOOC, and what has later become known as a connectivist MOOC, or cMOOC. The course enrolled over 2,000 students from around the world, and the learning was largely community-driven, allowing students to access a large array of open content, and participate in a wide variety of learning activities mediated through

technology. The course relied heavily on open technologies such as Moodle, RSS feeds, blogs, discussion forums and other collaborative synchronous and asynchronous tools [7].

The impact of MOOCs on higher education, and education in general, is still very difficult to measure. The number of stories about MOOCs in popular media, such as the New York Times, often presents MOOCs as an innovation in education. From a pedagogical perspective, most MOOCs rely on a traditional instructivist model, often relying on heavy use of video to convey content in a single direction, from instructor to student [8]. *Connectivism and Connective Knowledge* went largely unnoticed by the general public. It was not until Stanford's *Introduction to Artificial Intelligence,* launched in late 2011, that the public took notice of MOOCs, primarily because the course enrollment reached over 50,000 students within weeks. The striking difference was that the Stanford MOOC, later categorized as an xMOOC, was designed following a more instructvist approach, focusing primarily on one-way content delivery, compared to the Siemens and Downes cMOOC focused on collaboration. Classifying MOOCs as either a cMOOC or an xMOOCs might be somewhat misleading as many MOOCs contain elements of both. These terms might make more sense as two ends of a scale, where, depending on the design of a MOOC, it might lean towards one end of the scale or the other.

Measuring the impact of MOOCs on higher education is challenging. When educators look for metrics to measure the success of a course, we typically rely on metrics that are decades, if not centuries, old. Measures often used include persistence, such as examining the number of students that persist through a course to completion. Another metric includes success, often measured in education by those students that completed a course with a "C" or better. As MOOCs are an emerging field, educators are relying on traditional metrics to try and apply to MOOCs, even though this developing course format shares very little with what many view as 'traditional' higher education. New frameworks, such as the Distributed Intelligence Framework [9], might lead to better methods to assess the value of MOOCs for learners. For instance, this framework takes into consideration a learner's intentions, where in a MOOC environment not every student intends to finish the course. With this in mind, measuring MOOCs primarily by those that complete the cohort part of the course seems somewhat irrelevant. The forums of MOOCs represent a possible focal point for learners, providing a venue for tens of thousands of individuals to share ideas and insights around a common topic. In terms of learner intent, some learners might be motivated solely by the availability of thousands of peers in a single community, and have no intentions to complete any of the course assignments. While this appears to be a plausible reason to enroll in a MOOC, very little is known about how forums with up to 100,000 students provide value to learners.

To date, most research examining MOOC forums focus on the frequency of use and student responses to survey questions about the experience of using MOOC forums. One study, examining a MOOC offered by MIT, found that the surveys were the most frequently used

resource in the course, more so than lecture videos and homework assignments [10]. Another study found that in a cMOOC, the openness of the forums frequently led to negative experiences, as students felt overwhelmed by the number of posts and comments, and also discouraged to engage in the forums due to trolls (other forum posters that intentionally try to start arguments or upset other forum participants) [11]. While these studies are helpful, they both represent a sample of MOOC students willing to participate in a survey or interview. Due to the sheer volume of posts and comments on MOOC forums, manually reading, coding and analyzing these data is a daunting challenge.

2.2 Educational Data Mining

Educational Data Mining is an emerging area of research that employs data mining/machine learning algorithms to educational data in order to discover novel, previously unknown insights about how students learn [12]. The heterogeneity of educational data (e.g., student survey data, textual data from online educational environments, etc.) make data mining algorithm selection and applicability of extreme importance [13]. Data mining algorithms can be partitioned primarily into *unsupervised learning* and *supervised learning. Unsupervised learning* techniques such as clustering (e.g., X-means clustering) aim to discover natural patterns in an unlabeled data set [14]. For example, researchers in education may be interested in discovering the cluster of students that share similarity in learning styles (visual, textual, etc.), given a set of demographic or performance attributes. *Supervised learning* on the other hand aims to predict a class/output variable, given a set of mutually exclusive attributes. For example, researchers in the education domain may be interested in predicting student performance in MOOCs, given a set of different teaching styles (video lectures, text based lectures, etc.). Together, both *unsupervised* and *supervised learning* provide researchers with a wide array of data mining techniques that can be employed, given the research task of interest.

Researchers in the educational data mining community are successfully advancing student learning through innovative uses of data mining/machine learning algorithms. For example, Kelly and Tangney propose a data mining driven system that predicts students' learning styles based on a Naïve Bayesian machine learning model [15]. Minaei-Bidgoli *et al.* use features extracted from students' web logged data to predict their course performance [16]. Perera *et al.* focus on mining educational data in order to develop a better understanding of group behavior in online virtual environments [17]. One of the challenges in virtual education environments, compared to physical brick and mortar environments is the absence of direct student-teacher interaction during classroom instruction. Unlike student-teacher interactions in physical brick and mortar environment where students' facial expression/body language can communicate interest/disinterest in a course topic[18], educators in a virtual learning environment must rely primarily on textual information provided by students. Therefore mining student sentiment in virtual environments could serve as a critical dimension of educational data mining that may inform educators about how students learn over time.

2.3 Text Mining and Sentiment Analysis

Understanding what people think and how they feel has broad impact in fields ranging from marketing to psychology [19]. The advent of large scale textual data, generated through online social media platforms such as Twitter and Facebook, is providing researchers with rich sources of opinions expressed by users of these platforms. Opinion mining is an emerging research domain and has demonstrated tangible real life benefits. For example, researchers have mined the large scale textual data generated through Twitter to predict real life events in a wide range of applications such as financial stock markets to healthcare [20,21]. Tuarob and Tucker have quantified customer sentiment, expressed through social media sites, to predict product demand and preferences over time [22]. Hu and Liu propose techniques for mining customer opinions (positive or negative) in online product review sites [23]. Narayanan *et al.* propose a methodology to determine whether opinions expressed on different topics in a conditional sentence are positive, negative or neutral [24].

The data mining algorithms typically used in opinion mining include both *unsupervised learning* and *supervised learning* techniques. Given a vector containing textual data (e.g., a user's Tweet or Facebook comment), natural language processing algorithms such as Latent Semantic Analysis (LSA) [25]or Latent Dirichlet Allocation (LDA) [26] can quantify the similarity among textual documents (e.g. similarity between two customer reviews). The sentiments/opinions of a user can be quantified by analyzing the individual text expressed by a user and assigning a sentiment score (positive, negative or neutral), based on how these words are used in human communication (e.g., the word *love* being a positive word while the word *hate* being a negative word). The authors of this work aim to understand how student sentiments (expressed textually) in MOOCs impact overall student performance over time. Such valuable insights will enable educators to develop intervention mechanisms aimed at increasing student interest and performance in MOOCs. Students will also benefit from this knowledge by understanding how textual content (primarily expressed through online discussion forums) can propagate throughout student networks (e.g., project groups) and impact performance.

3. Methodology

The methodology presented in Figure 1 mines student-generated textual data (e.g., online discussion forums) existing in MOOCs in order to quantify their impact on student performance and learning outcomes. In this work, student performance is limited to grades attained in course homework assignments, quizzes and examinations. The methodology in Figure 1 starts with the acquisition of MOOC data (text), followed by the organization and storage of this data in a traditional SQL database. Sentiment analysis algorithms are them employed on the textual data in order to quantify the aggregate student sentiment pertaining to each assignment. The methodology ends with a statistical analysis that quantifies the correlation between student sentiment and student performance. Temporal patterns in student sentiment, in relation to student

performance are also investigated for a deeper understanding in how student sentiments evolve over time.

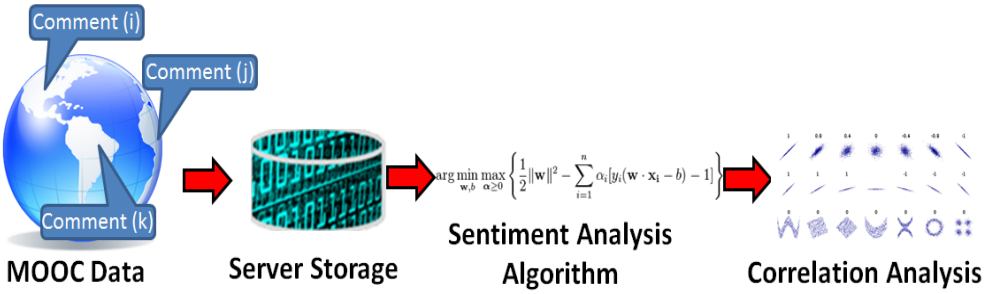


Figure 1: Mining MOOC Text Data for Students' Sentiments

3.1 Data Acquisition

Student data, generated through Massive Open Online Courses (MOOCs) is primarily textual in nature and is expressed in open discussion forums labeled as *posts* (a new discussion thread initiated by a student/instructor) or *comments* (a response to a *post* by a student/instructor). For each student that participates in an online discussion topic (*post*), their unique ID, in addition to the time of their post is recorded. For example, for the *Introduction to Art: Concepts and Techniques* MOOC data analyzed in the experimental study section of this work, a comment from one of the students is of the form:

```
[Hi:) I really like your photos, very delicate and clean :)<br
/>]
```

The student comment above is combined with both textual data and emoticons (e.g., :)), highlighting the challenges of quantifying sentiment in a large corpus of textual data such as that found in MOOCs. The sentiment analysis algorithm employed in this work can quantify both textual sentiments (e.g., the word "love" being classified as a positive sentiment) and emoticon based sentiments (e.g., :) being classified as a positive sentiment). In addition to this, the data has to also be preprocessed for HTML commands such as <*br*>, which represents a line break command. In this work, the raw data is preprocessed to remove HTML commands, prior to being stored and mined for student sentiments.

3.2 Server Storage

The student MOOC data (primarily comprising of texts and emoticons) is stored in a traditional SQL database for efficient querying. Each student has a unique student ID and while the actual demographic information (name, age, etc.) is typically not available in MOOCs, the student ID enables researchers to quantify what posts relate to a given student ID. Actual demographic information may be specifically solicited by the instructor, however the authenticity of that data cannot be guaranteed. Nevertheless, the availability of a unique ID associated with each student,

enables researchers to organize the data on the server in such a manner that all *posts* or *comments* by a particular user ID can be aggregated and returned with standard SQL commands. The data stored on the server is organize in tables, with each column in a table, representing an attribute (e.g., time) of the MOOC data.

3.3 Sentiment Analysis Algorithm

The sentiment analysis algorithm employed in this work is based on the word-sentiment lexicon proposed in [27,28] , enabling researchers to take into account both student sentiments relating to specific words (e.g., love) or sentiments relating to specific emoticons (e.g., :)). A Sentiment Orientation (SO) refers to the *polarity* and *strength* of words, phrases, or texts, where *polarity* refers the positive, negative or neutral characteristics of a student sentiment and *strength* refers to magnitude of that sentiment. For each student *post* or *comment* found in the MOOC data, each word is automatically mapped to the positive or negative emotion value using the following scales [27]:

[no positive emotion or energy] 1– 2 – 3 – 4 – 5 [very strong positive emotion]
[no negative emotion] 1– 2 – 3 – 4 – 5 [very strong negative emotion]

Sentiments, expressed within student posts/comments are weighted based on multiple factors such as the emoticon used to emphasize a textual sentiment (e.g., "I am very happy about my quiz grade :)", a negative word that alter potentially positive sentiments (e.g., "I am *not* very happy about my quiz grade), etc. Since a single student post can express multiple sentiments, sentiment score bounds have theoretical bounds of -∞ to ∞. Quantifying sentiments over time enables researchers to understand the temporal variations in student opinions towards certain course topics, group discussions or instructor performance.

3.4 Correlation Analysis

In this work, researchers aim to understand the correlation between student sentiments, expressed through MOOC posts/comments and student performance using the following equation:

$$\boldsymbol{Corr} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad \textbf{(1)}$$

Where,
X_i**:** represents the student sentiment (individual or aggregate) for a given assessment (e.g., quiz 1)
$\bar{X}$: represents the average student sentiment for all assessments being investigated
Y_i**:** represents the student grade (individual or aggregated) for a given assessment (e.g., 90/100)
$\bar{Y}$: represents the average student grade for all assessments being investigated

N: represents the total number of samples being analyzed (e.g., total number of quizzes in the semester

4. Experimental Study

4.1 Coursera Art MOOC Course Description

Coursera Art MOOC, *Introduction to Art: Concepts and Techniques* was based on Art 10: *Introduction to Visual Studies* taught by Anna Divinsky at Penn State University (PSU). Much of its presentation, formatting, rubric, announcements, assignments and quizzes were specifically tailored to Coursera's platform and audience.

The Art MOOC was a 7-week course designed for learners without any previous art knowledge or experience. It focused on giving the students a taste of various art forms, artists, and artworks – each chapter introducing the students to a different art movement, style, and discipline. The overarching goal for this Art MOOC was to expose the students to new art concepts, encourage art making and experimenting, as well as careful consideration and awareness of the conceptual aspect of each assignment. Students were asked to provide a short artist statement along with each assignment submission where they explained their concept and process, thus articulating their ideas through writing and expressing themselves creatively. Another space for self-expression was the discussion forums where students were able to post and answer questions and most importantly participate in class discussions.

The course was rich with various forms of content such as text with images, artist feature videos, quizzes, assignments, artwork examples created by PSU students, and a wide range of instructional videos that addressed art techniques and materials as well as creative approaches to each assignment. The Artist Feature Videos included in-line quiz questions, or self-checks that enabled the students to check their understanding as they progressed through the videos.

Peer-evaluation was a crucial part of the course that allowed the students to share their artwork with one another, evaluate it using a simple rubric, and then provide one another with personal, constructive feedback that would help each grow and improve as they progressed through the course. For the evaluation process, the learners were automatically matched with two classmates, but could choose to evaluate more. Many students enjoyed this process, because they were able to share their ideas and suggestions.

In the beginning of the course, the learners could choose from two different tracks "studio" or "non-studio", providing them with an option of having a hands-on art experience or competing just the readings and quizzes. Students received a Certificate of Accomplishment after completing five quizzes with an average of 70%. Those who also submitted 2 assignments were awarded a Statement of Accomplishment with Distinction. It was interesting to see that many

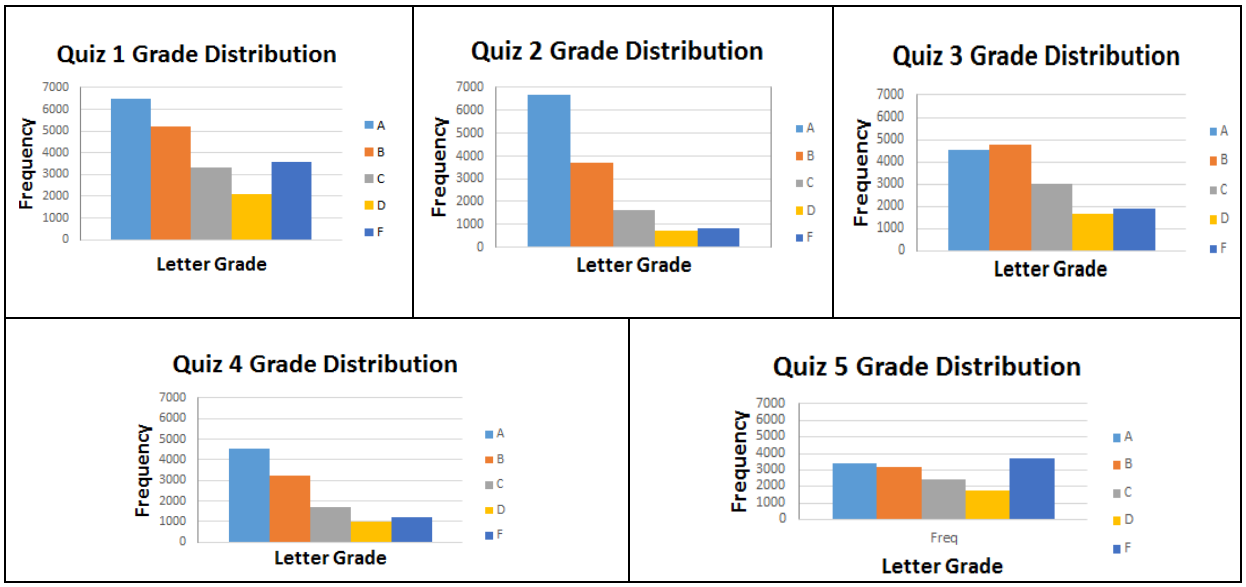
students, who in the beginning were not planning to create artwork, enjoyed the art making aspect of the course so much that they wanted to submit their assignments and evaluate their peers' work. Section 5 discusses the insights discovered by mining the Art MOOC data for student sentiments, expressed textually through discussion forums.

5. Results and Discussion

5.1 Coursera Art MOOC Grade Distribution

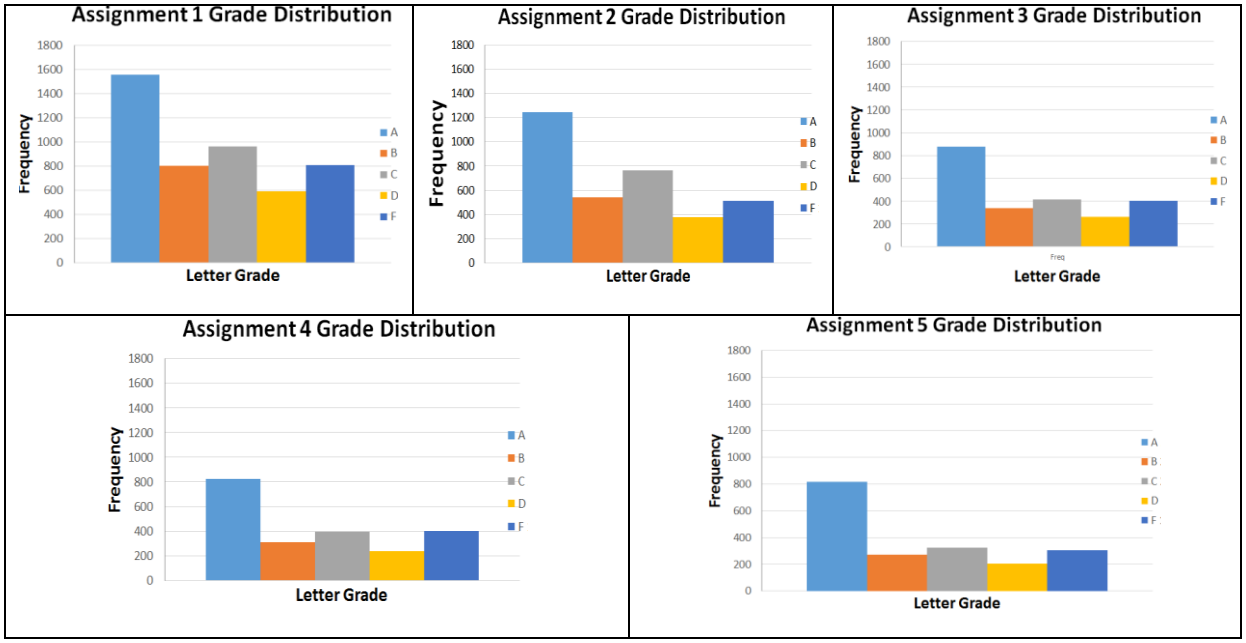
Tables 1 and 2 provide an overview of the distribution of grades for the entire duration of the Art MOOC. Table 1 provides the grade distribution for quizzes, while Table 2 provides a distribution of the assignment grades. Together, these modes of assessing student performance represent the output variable to be used in evaluating the correlation between students' sentiments expressed in posts/comments and student performance. It is interesting to note that in both the quiz assessments (Table 1) and assignment assessments (Table 2), there appears to be a relative decrease in the frequency of A's earned in the class, possibly alluding to the increase in course material difficulty over time, decrease in student interest, or some other latent factor to be investigated.

Table 1: Distribution of Quiz Grades for the Art MOOC Course



Another interesting observation between the quiz (Table 1) and assignments (Table 2) distributions is the population of students that complete each assessment. For the quiz distributions in Table 1, a significant number of students complete the quiz assessment (21,702), compared to 4,732 students who completed the assignment 1 assessment. A similar pattern can be observed in each of the 5 quiz completion rates, compared to the 5 assignment completion rates.

Table 2: Distribution of Assignment Grades for the Art MOOC Course



5.2 Coursera Art MOOC Sentiment Analysis

The sentiment analysis of the Coursera Art MOOC begins with a plot of students' sentiment values over time. Figure 2 illustrates the difference between student sentiments relating to *posts* (the start of a new topic of interest) and *comments* (textual responses relating to a given post). The timeline in Figure 2 is represented in Computer Epoch time, where 1369702406, represents Monday, May 27 2013 20:53:26 GMT-0400 (Eastern Daylight Time) in a standard date format.

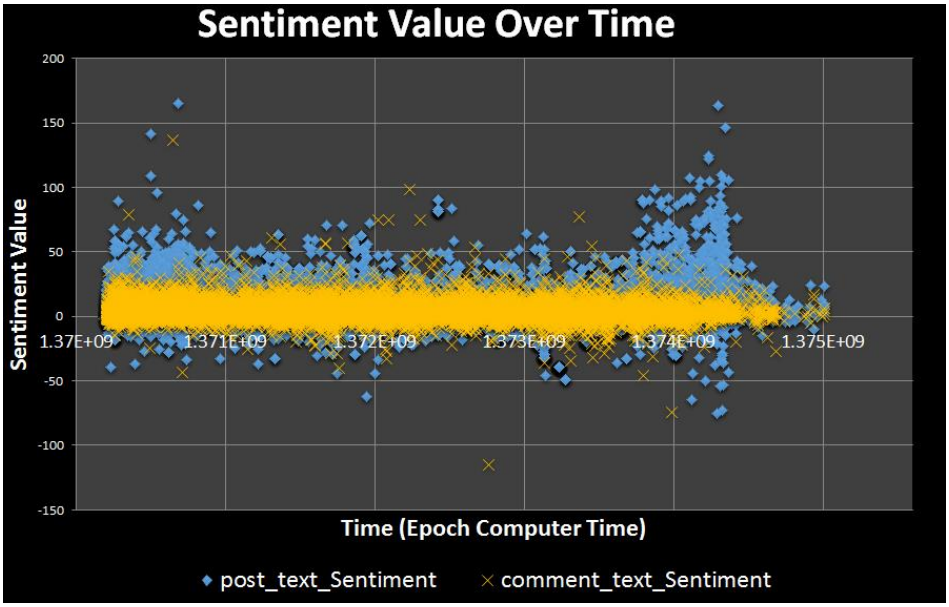


Figure 2: Quantified Student Sentiment Values over Time

Figure 2 indicates that the sentiment expressed in student *posts* over time demonstrates greater variance, compared to student *comments*, with average sentiment values of 6.03 and 3.90

respectively. Furthermore, Figure 2 indicates that the variance in sentiment value of student *posts* diverges towards the end of the course offering timeline, while the student *comments* reveal an opposite trend. The researchers of this work postulate that a possible reason for differences in student sentiment between *posts* and *comments* (especially towards the end of the semester) is that, as students stress out about completing assignments, etc., their sentiments when initially discussing a course topic through a *post* may be more intense than the responses to that post in the form of comments/discussions. Another possible reason for this is the large number of posts in general, happening so quickly that many often get lost in the forum and receive no comments. However, further research is needed to investigate these phenomena which are a topic for future work.

5.3 Coursera Art MOOC Correlation Analysis

The correlation analysis performed on the Art MOOC textual data reveal interesting findings relating to student sentiments (averages) and performance (averages) for given assessments. For quizzes (Figure 3), there is a slightly positive correlation of 0.320 suggesting that student sentiments expressed in the discussion forums (including both posts and comments) relating to quizzes may not be a good indicator of student performance. With regards to the assignments however, there is a stronger (negative) correlation between student sentiments expressed in discussion forums relating to assignments and actual average assignment scores, with a correlation value of -0.820. Researchers were surprised by the negative correlation between student sentiment and assignment scores.

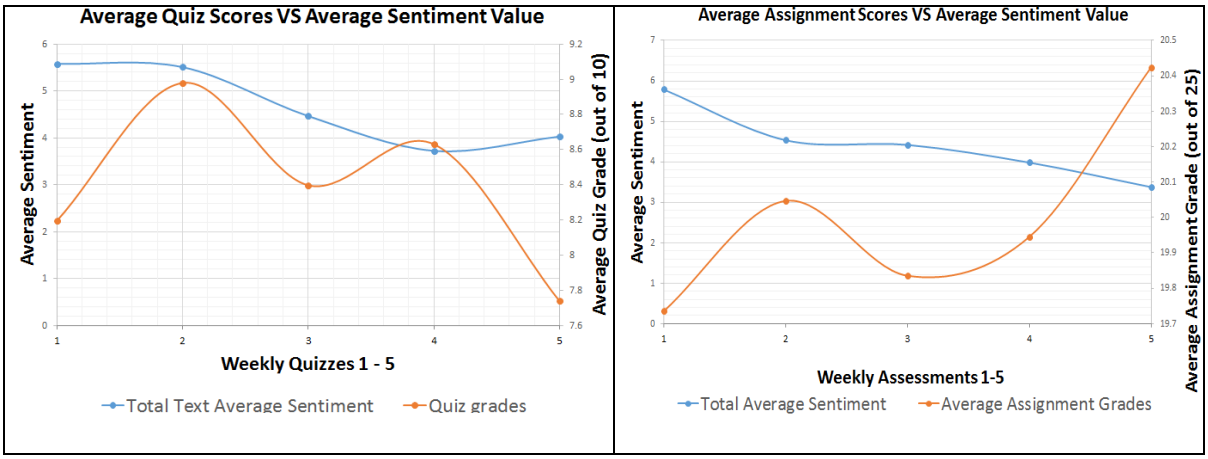


Figure 3: Correlation Analysis of Student Sentiment Values and Student Performance

That is, as the assignment scores increased, the sentiment in the discussion forums pertaining to assignments actually decreased (i.e., student expressed more negative emotions). While there are many reasons that could explain this, the researchers postulate that this decreased sentiment score could be due to the increased expectation of students in terms of the quality of feedback

that they demanded, which in turn helped them perform better on the subsequent assignments. The nature of peer assessment in MOOCs, where novices are grading novices, also might influence these results. Due to the fact that students know peers are grading them, perhaps students are more likely to publicly voice negative comments about a grade. Future work aims to test several hypotheses that could help explain these phenomena.

6. Conclusion and Path Forward

The objective of this research is to mine student-generated textual data (e.g., online discussion forums) existing in MOOCs in order to quantify their impact on student performance and learning outcomes. Two aspects of student performance were investigated; quizzes and homework assignments. Initial research findings reveal that student sentiments were slightly (positively) correlated with quiz performance (0.320), while more strongly (negatively) correlated with homework assignments (-0.820). Future work in MOOC sentiment analysis aims to advance beyond quantifying the correlations between student sentiment and performance, towards a deep understanding of why these correlations exist in the first place. In future work, the authors will also investigate the relevance of other MOOC features (such as time spent in online forums, geographic dispersion of students, etc.) in predicting students' performance in MOOCs. Furthermore, the authors of this work aim to repeat this study in future offerings of the Coursera Art MOOC in order to compare the research findings across different time periods, student demographics, etc. The authors of this work are part of a cohort of data scientists at Penn State University and are working towards analyzing the data generated in a wide variety of MOOCs recently launched by Penn State University, in order to investigate the research findings that are common across a wide range of MOOCs. Another area of potential research expansion is to investigate hybrid courses that utilize both brick and mortar and online modes of education delivery.

References

1. Internet users (per 100 people) | Data | Table. Available at: http://data.worldbank.org/indicator/IT.NET.USER.P2. Accessed December 12, 2013.
2. MIT OpenCourseWare | Free Online Course Materials. Available at: http://ocw.mit.edu/index.htm. Accessed December 12, 2013.
3. Online Schools, Classes, Degree Programs - University of Phoenix. Available at: http://www.phoenix.edu/. Accessed December 12, 2013.
4. Clow D. MOOCs and the funnel of participation. In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. LAK '13. New York, NY, USA: ACM; 2013:185–189. doi:10.1145/2460296.2460332.
5. Green K. Massive Open Online Courses (MOOCs) and Other Digital Initiatives. *J Collect Bargain Acad*. 2013;(8). Available at: http://thekeep.eiu.edu/jcba/vol0/iss8/10.
6. Yuan L, Powell S. MOOCs and Open Education: Implications for Higher Education. 2013. Available at: http://publications.cetis.ac.uk/2013/667. Accessed December 18, 2013.

7. Siemen. Massive Open Online Courses: Innovation in Education? In: *Perspectives on Open and Distance Learning: Open Educational Resources: Innovation, Research and Practice*. COL, Athabasca University; 2013:5–15.
8. Daniel J. Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *J Interacti*. 2012;3.
9. Grover S, Franz P, Schneider E, Pea RD. Distributed Intelligence Framework for the Design and Evaluation of MOOCs. In: Madison, WI; 2013.
10. Breslow L, Pritchard DE, DeBoer J, Stump GS, Ho AD, Seaton DT. Studying learning in the worldwide classroom:Research into edX's first MOOC. *Res Pract Assess*. 2013;8:13–25.
11. Mackness J, Mak SFJ, Williams R. The Ideals and Reality of Participating in a MOOC. In: *Seventh International Conference on Networked Learning*. Aalborg, Denmark; 2010. Available at: http://www.lancaster.ac.uk/fss/organisations/netlc/past/nlc2010/abstracts/Mackness.html.
12. Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. *Expert Syst Appl*. 2007;33(1):135–146. doi:10.1016/j.eswa.2006.04.005.
13. Romero C, Ventura S. Educational Data Mining: A Review of the State of the Art. *Syst Man Cybern Part C Appl Rev IEEE Trans On*. 2010;40(6):601 –618. doi:10.1109/TSMCC.2010.2053532.
14. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. Elsevier; 2011.
15. Kelly D, Tangney B. *"First Aid for You": Getting to know your Learning Style using Machine Learning*.
16. Minaei-Bidgoli B, Kashy DA, Kortmeyer G, Punch WF. Predicting student performance: an application of data mining methods with an educational Web-based system. In: *Frontiers in Education, 2003. FIE 2003 33rd Annual*.Vol 1.; 2003:T2A–13–18 Vol.1. doi:10.1109/FIE.2003.1263284.
17. Perera D, Kay J, Koprinska I, Yacef K, Zaïane OR. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Trans Knowl Data Eng*. 2009;21(6):759–772. doi:10.1109/TKDE.2008.138.
18. D'mello S, Graesser A. Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments. In: *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*. Amsterdam, The Netherlands, The Netherlands: IOS Press; 2007:161–168. Available at: http://dl.acm.org/citation.cfm?id=1563601.1563631. Accessed December 12, 2013.
19. Pang B, Lee L. Opinion Mining and Sentiment Analysis. *Found Trends Inf Retr*. 2008;2(1-2):1–135. doi:10.1561/1500000011.
20. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *J Comput Sci*. 2011;2(1):1–8. doi:10.1016/j.jocs.2010.12.007.
21. Salathé M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Comput Biol*. 2011;7(10):e1002199. doi:10.1371/journal.pcbi.1002199.
22. Tuarob S, Tucker CS. Fad or Here to Stay: Predicting Product Market Adoption and Longevity Using Large Scale, Social Media Data. In: *Proceedings of the 2013 ASME IDETC/CIE, DETC2013-12661*. Portland, Oregon: ASME.
23. Hu M, Liu B. Mining Opinion Features in Customer Reviews. In: Mcguinness DL, Ferguson G, Mcguinness DL, Ferguson G, eds. AAAI Press / The MIT Press; 2004:755–760. Available at: http://dblp.uni-trier.de/rec/bibtex/conf/aaai/HuL04.
24. Narayanan R, Liu B, Choudhary A. Sentiment Analysis of Conditional Sentences. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009:180–189. Available at: http://dl.acm.org/citation.cfm?id=1699510.1699534. Accessed December 22, 2013.
25. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci*. 1990;41(6):391–407.
26. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993–1022.
27. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment strength detection in short informal text. *J Am Soc Inf Sci Technol*. 2010;61(12):2544–2558. doi:10.1002/asi.21416.
28. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based Methods for Sentiment Analysis. *Comput Linguist*. 2011;37(2):267–307. doi:10.1162/COLI_a_00049.