Agricultural Journal 13 (5-6): 107-113, 2018
ISSN: 1816-9155

Learning Analytics in Massive Open Online Courses (MOOCS)-Assessing Participant's Performance

Rabie Moussa Houssin and Mohammed Nazeh Abdulwahid
Limkokwing University of Creative Technology, Cyberjaya, Selangor, Malaysia

Abstract: This study investigated the gap of how the differences of a personal profile, subjects and culture may moderate the learner's performance in Massive Open Online Course (MOOC). One of the learning analytics in MOOC is assessing and predicting learner's performance, dropout and completion. The dataset consists of 641,138 observations of 20 variables. Based on literature review and analysis of the dataset variables twelve variables are chosen. For measurement purposes, two variables "grade" and "certification" which account for learning outcomes are available to the researcher. Five hypothesis are examining direct relations while another four set of relations are examining the moderation affect. The model summary measures show that this accounts for only 4.9% of variance in the spelling grades. Interaction log forum log and chapters log associated hypothesis are rejected. Interaction days shows a significant standardize regression coefficients value (Beta = 0.219) in the model and the hypothesis is accepted. Video log shows a significant standardize regression coefficients value (Beta = 0.217) in the model where hypothesizes are discussed and results are figured out.

Key words: MOOC, interaction log, interaction days, video interaction rate, chapters interaction rate, forum interaction rate, student's performance

INTRODUCTION

In short, Massive Open Online Course (MOOC) is that can be defined as web-based online courses which is different from the classic e-Learning systems. The meaning of MOOC is; massive: unlike traditional online education systems it aims large audiences worldwide. Open: there is no prerequisite or formal procedure for participation in courses. They are open to the worldwide because they are mostly free of charge or only certain modules are paid such as certification. Online: all educational contents and educational activities such as assignments and exams are online and accessible via. the internet as web-based. Course: the course content includes the following criteria compared to e-Learning systems. Instructional: with the purpose of achieving the highest instructional level, presentations, documents and additional resources are given in addition to 8-10 min video contents appropriate to the pedagogical format which are generally divided into modules. Certification: they provide certified job-ready courses along with technology companies. Learning activities: especially the courses in the field of informatics include offline or online interactive encoding activities to increase the instructional level. In addition to this, activities such as assignment, quizzes, final exam and final project are also conducted depending on a particular calendar.

Learning network: they provide the formation of e-Learning networks in which activities such as problem solving, mentor support and discussion are conducted for each course for the development of courses and for trainees to conduct learning activities together.

One of the learning analytics in MOOC is assessing and predicting learner's performance, dropout and completion based on differences between semi-structured data that collected from discussion boards, grades, attendance and other metrics. In the past, all the studies verify a directed regression relation between learner's system interaction and their performance, completion or dropout there. In the past problem is no generally agreed-upon data model for student interactions. Thus, analysis tools must be tailored to each system's particular data structure, reducing their interoperability and increasing development costs. Some e-Learning standards designed for content interoperability include data models for gathering. Student performance information which we link to how two e-Learning standards IEEE standard for learning technology and experience API define their data models. This study will explore the impact of student's activities and their performance based on their personal profiles (gender, age, education and country) may moderate the learner's performance in MOOC. The study is going to explore research problems to identify the learner's activity logs (fields). The aim is to build a

Corresponding Author: Rabie Moussa Houssin, Limkokwing University of Creative Technology, Cyberjaya, Selangor, Malaysia

method that evaluates the model of learner's performance predictors as an assistant aid for better decision making (De. Barba, 2016).

Literature review: The studydomain belongs to MOOCs learning analytics as a main concept but the conceptual framework includes big data and MOOCs as well. The most frequent definitions among past literature identify big data as a huge dataset characterized with a combination of three aspects-the three Vs-Volume, Velocity and Variety. Massive Open Online Course (MOOC) can be defined as web-based online courses are different from the classic e-Learning systems by the following characteristics: massive, open, online instructional course, certification based and enriched with learning activities. There are two forms distinct, the Canadian MOOCs came to be called c MOOCs (for "connectivity" or many-to-many network) and the ones from Coursera, ed X or Udacity came to be called "x MOOCs" (for "exponential" or one-to-many network) (Barak *et al*., 2016). The specific application of Data Science (DS) in the education field is known as Educational Data Science (EDS) which works with data gathered from educational environments/settings to solve educational problems. There are a variety of technical methods to deal with big data, whose key is to synthetically make use of analysis techniques to provide learners for better study environment and learning advice such methods are: statistical analysis, Social Network Analysis (SNA), content analysis, data mining and data visualization. Next sections are going to discuss the methodology which include, dataset description, proposed model and analysis techniques. Student performance information which we link to how two e-Learning standards IEEE standard for learning technology and experience API define their data models. It is from this study will explore the impact of student's activities on the studen't performance and how the differences of a personal profile (gender, age, education and country) may moderate the learner's performance in MOOC. The study is going to explore research problems, identify the learner's activity logs (fields), build a method that evaluates the model of learner's performance predictors and controllers in MOOC to examine the new model for predicting student's performance. We identify the advantages of using these e-Learning standards from the point of view of learning analytics. However, scholars recommended for further research to investigate the moderation effect of extra indirect factors. We will in this study, explore the gap of how the differences of a personal profile, subjects and culture may moderate the learner's performance in MOOC. The research gap is allocated in the regression analysis techniques of MOOC.

Fig. 1: Data complexity

Big data concept: The term Big Data "BD" is often used to describe datasets that have grown in size well beyond exabytes and zettabytes. These datasets reach a point where the ability to capture, manage and process such, items within a reasonable amount of time, cannot be achieved with commonly used software tools (Abubakar *et al*., 2014). Initially the idea of big data was addressed was in June 2006. It had been referred to as vast levels of data which in turn traditional methods could not process because of intricacy and volume (Al-Fawaz, 2012).

Big data definition: Recent research describe big data simply because ways to course of action huge (from terabytes to exabytes) and multiparty (varies from Messfuhler to sociable media) data series that demand storage, process, research and creation methods (Chatti *et al*., 2012). The most frequent definitions among past literature identify big data as a combination of three aspects-the three vs. Volume, Velocity and Variety (Chang *et al*., 2014). These three main dimensions of big data are explained in Fig. 1.

Volume: It is the dimension most associated to big data and refers to data's size (Chang *et al*., 2014). However, it does not necessary has to do with bytes, terabytes and petabytes, since, several companies quantify data in terms of time (e.g., 7 years of data available for risk and legal analysis) (Arora and Vermeylen, 2013). In accordance to storage is not today the most complex issue about big data, although, the ability to process or move data from one storage facility to other might take large amounts of time. Moreover, now a days it is not viable to transport large amounts of data from one data set to another, creating the need for the analysis to be processed at distance and therefore, requiring higher analytics capacity.

Velocity: This implicates the processing capacity at real or almost real-time speed (Barocas and Nisscnbaum, 2014). Data is created at an accelerated pace and firms are currently collecting and storing it at an outstanding pace (Arora and Vermeylen, 2013). Nonetheless, several applications are only achievable if the processing capacity is doable at a real-time basis, diminishing the time between data capture and the moment it becomes accessible and it is analysed. Moreover, according to Arora and Vermeylen (2013) some activities require real time processing capacity in order to enrich firm's efficiency. Since, traditional methods do not have the capacity to obtain results in real or almost real time speed, big data analytics are needed to provide a proper response to these challenges. Additionally the combination of the first V Volume with the second, V Velocity increases the need for high processing ability as high volumes of data which require processing ability to be at real or almost real time increase the need for higher and more complex processing and data mining applications.

Variety: Being the third vertex of the big data Vs. triangle, variety refers to the wide spectrum of data formats (Barocas and Nissenbaum, 2014). Structured, semistructured and unstructured data, audio, image, text, social networking data needing contextual information compose now a days the mixture of data available. Despite the wide variety of data already being collected by organizations, only recently are they beginning to analyze data from different streams as one large. In addition, both volume, velocity and variety tend to boost each others. Another vertex was added to big data's definition by (Barocas and Nissenbaum, 2014). The fourth "V" Veracity represents the level of data's reliability, since, some contents are not considered to be a reliable source of information.

Nonetheless, big data is just data, therefore, it does not represent any competitive advantage without analysis on top. As Aiden and Michel (2013) said: "Big data is not new but the effective analytical leveraging of big data is". Big data analytics consists of techniques able to process big data in all its three dimensions (Aiden and Michel, 2013; Schroeck, 2012). The next subsection big data analytics-presents an analysis about the rise of big data analytics its fields of usage and forms of applicability.

Big data analytics: Big data analytics includes the techniques to power big data benefits (Aiden e and Michel, 2013; Barocas and Nissenbaum, 2014). Despite users being conscious of the large info creation, the idea of big data analytics continues to be likely to be unfamiliar (Aiden and Michel, 2013; Barocas and Nissenbaum, 2014). According to a study done by Russom, 65% of individuals understand the implications of big data analytics but usually do not acknowledge the name and 7% of these have not found out about the idea itself. Agencies perceive big data analytics as a combined band of ways to analyze today's huge amounts of data, providing fresh insights to Ieverage firm's performance (Adler-Milstein and Jha, 2013).

Furthermore, according to big data analytics exists today in several companies not only as big data analytic's providers but also as customers, purchasing big data analytics insights to improve their businesses. According to a survey conducted by Alla (2013), the five most common big data techniques used by organizations are reporting, data mining, data visualization, predictive modelling and optimization services. Such big data techniques can be used by organizations for different purposes: to achieve a higher level of efficiency in their processes throughout reporting and optimization techniques or a higher segmentation of customers with the use of data mining technologies and a better and faster response to their customers changes. However, and despite big data analytic's wide spectrum of capabilities, only 6% of organizations are currently using big data techniques (Chang *et al*., 2013). Moreover, according to, the researchers 24% did not use such techniques neither plans of using them in the near future whereas 47% are beginning to plan and conducting a roadmap to incorporate such capabilities in their strategy. Finally, around 22% of organizations begin to prove the business value of big data techniques as well as perform an assessment of their technologies and skills. Despite big data analytic's wide spectrum of capabilities and sectors of activity, several changes are faced. Being a relatively new stream of technology it is important for both organizations and final users to understand big data opportunities and how it can leverage and influence firm and customer's benefits (Boyd and Crawford, 2012; Manyika, 2011).

Moreover, the learning curve to adopt big data technologies is fast but requires effort to keep track of the most recent technologies developed (Chang *et al*., 2014). As a matter of fact, previous studies state that only one third of companies have the goal to change its analysis throughout big data analytics technologies whereas roughly 47% will keep their analytics without big data techniques (Adler-Milstein and Jha, 2013). Additionally, usually big data technologies require IT specialists to perform and develop analysis that provide useful insights to influence companie's strategy. According to, the USA alone will face a shortage of 190.000-140.000 IT specialists in the short-term. This creates a barrier to the spread of this type of technologies across the entire market. Another challenge big data

analytics faces is its benefit's lack of perception by end users in which value creation by companies might not be followed by its correspondent value capturing. Finally, the most popular challenge associated to big data is related to end-user's privacy concerns (Boyd and Crawford, 2012).

MATERIALS AND METHODS

As discussed previously, both institutions and instructors are eager to understand how students are engaging with MOOCs. To explore the determinants of learner's performance in MOOC, to apply user profile for improving the prediction technique that will help or associate improving the performance of e-Larning using (MOOC) and to analyse the performance of e-Learning using computational regression model based on our proposed user profile.

In this study, we will to uses a secondary data set of Harvard X-MITx Person-Course Dataset AY2013. The data collection was conducted among course participants from the whole world. Data analysis is using multiple regression and data visualization by using SPSS using a quantitative design because numerical analysis nature of the data representation and analysis in order to assessing student's performance antecedents and controllers of MOOC-based systems. Quantitative research produces numerical data or information interpreted as numbers where the study aims to discover the question through numerical proof. Quantitative techniques normally use the scientific research steps, starting by producing a hypothesis to be tested. The suitable data-set is analyzed for analysis of: demographic, descriptive and relations. The findings were discussed in conjunction with the proposed hypothesis to achieve external validity of the proposed model. The hypothesis must be provable by statistical and mathematical measures and is the center that the whole study is focusing in. In this research, some objectives were answered by using quantitative approach. Statistical analysis based on descriptive measures, variance, covariance techniques are used to find out and propose the desired answers to the three objectives.

Regression analysis is undertaken to analyze the impact of previous educational attainment and engagement in the course on final grade and certification of registrants. Various regression models are fitted to understand these dynamics. The software used for descriptive statistics, regression analysis and graphical presentations is SPSS with the exception of some plots which are generated through other means. It should be noted that the data is already "cleaned" where the outliers (in this case individuals with unusually high activity in any course) have already been removed by the dataset provider . This serves as one of the key strengths of the data as the extreme values are expected not to impact the results unnecessarily, hence, yielding more robust estimates.

Hypothesis:

- H_1: interaction log have a positive impact of students performance of MOOC's participants
- H_2: interaction days have a positive impact of student's performance of MOOC's participants
- H_3: video interaction rate have a positive impact of students performance of MOOC's participants
- H_4: chapters interaction rate have a positive impact of students performance of MOOC's participants
- H_5: forum interaction rate have a positive impact of students performance of MOOC's participants
- H_6: gender of learner moderates the prediction model of student's performance of MOOC's
- H_7: age of learner moderates the prediction model of student's performance of MOOC's
- H_8: education level of learner moderates the prediction model of student's performance of MOOC's
- H_9: country of learner moderates the prediction model of student's performance of MOOC's

Collection and data analyses: Data Science (DS) is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data and analysis on data. The practice of data science in educational dataset MOOCs is of great interest to Data Science (DS) researchers and known as Educational Data Science (EDS). The specific application of Data Science (DS) in the education field is known as Educational Data Science (EDS) which works with data gathered from educational environments/settings to solve educational problems.

This study is emphasizing on the topic of learner performance in MOOCs based on secondary data of Harvard X-MITx Person-Course Dataset AY2013. The data collection was conducted among international participants, therefore, the scope of the study population is learners who join the associated coursed in 2013.

The study scope of interest is learner performance in MOOCs, any other types of online or traditional learning is not related to this study. In addition, the data is examining the learner performance and any other parties such as lecturer, tutors, management of MOOC system are out of the scope.

Learner performance in MOOCs is the main measurement and dependent variable to assess and only direct relations with the determinants were investigated. Moreover, the data are based on secondary data that collected automatically by the system (IS log). No data is collected directly from learners of any other party.

RESULTS AND DISCUSSION

For coming out with the study findings, first the data screening has been implemented for checking the raw data, identifying outliers and dealing with missing data and the result is shown in Table 1.

As shown in Table 1, the total cases were 641.138 cases, the filtered cases with complete field were 13678 cases and the final clean data after removing strange cases were 13232 cases. The average of total number of valid respondents is 13232 learners. Demographics section of the data set contains information related: age, gender, country and education. Male learners are 70% while female are 30%; The main age group is 25-34 with 43% followed by the age group 18-24 with 40%; The majority of learners by MOOC are holding bachelor degree with 35.5%; It logical to find that 23% of the learners are from USA, followed by India with 20%. It is clear that forum interaction is very week with a maximum count of 6 while interaction of the videos has a maximum of 390 per learner. Interaction log have a huge number because it is a log of all interactions including academic and nonacademic entries. The measures show that all relations are significant with $p<0.001$. While the correlation is significant but it seems that all the relations are in weak affect. The model summary measures show that the model is significant with $r = 0.222$ and the adjusted $R^2 = 0.049$ tells us that our model accounts for only 4.9% of variance in the spelling grades. Interaction log shows a significant standardize regression coefficients value (Beta = -0.341) in the model but the hypothesis is rejected. Interaction days shows a significant standardize regression coefficients value (Beta = 0.219) in the model and the hypothesis is accepted. Video log shows a significant standardize regression coefficients value (Beta = 0.217) in the model and the hypothesis is accepted. Chapters log shows a significant standardize regression coefficients value (Beta = -0.062) in the model and the hypothesis is rejected. Forum log shows a significant standardize regression coefficients value (Beta = -0.040) in the model and the hypothesis is rejected. The four learner's profile characteristics is found to be moderators in the prediction model of student performance.

As shown in Table 2, three hypothesis (H_1, H_4 and H_5) are rejected while six hypothesis (H_2, H_3, H_6, H_7, H_8 and H_9) are accepted as shown in Fig. 2.

Table 1: Data screening results

Description	Count
Data Full-Set	641.138
Data Complete-Fields	13678
Initial Cases for Analysis	13678
Univariate Screening	446
Valid Data-Set	13232

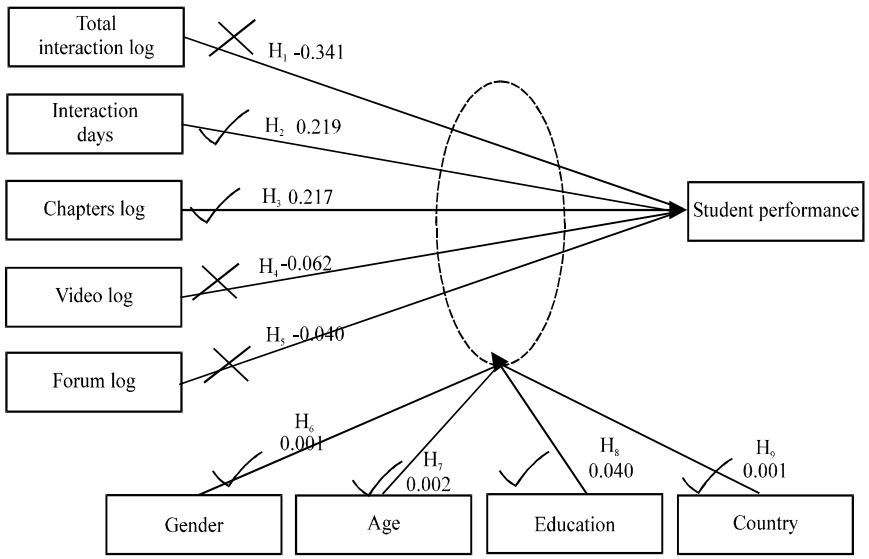

Fig. 2: Supported and not supported hypothesis in the proposed model

Table 2: Summarizes these results

Hypothesis	Status
H_1: interaction log have a positive impact of students performance of MOOC's participants	Rejected
H_2: interaction days have a positive impact of student's performance of MOOC's participants	Accepted
H_3: video interaction rate have a positive impact of students performance of MOOC's participants	Accepted
H_4: chapters interaction rate have a positive impact of students performance of MOOC's participants	Rejected
H_5: forum interaction rate have a positive impact of students performance of MOOC's participants	Rejected
H_6: gender of learner moderates the prediction model of student's performance of MOOC's	Accepted
H_7: age of learner moderates the prediction model of student's performance of MOOC's	Accepted
H_8: education level of learner moderates the prediction model of student's performance of MOOC's	Accepted
H_9: country of learner moderates the prediction model of student's performance of MOOC's	Accepted

CONCLUSION

The research domain belongs to MOOCs learning analytics as a main concept but the conceptual framework includes big data and MOOCs as well. The most frequent definitions among past literature identify big data as a huge dataset characterized with a combination of three aspects-the three Vs.Volume, Velocity and Variety: Massive Open Online Course (MOOC) can be defined as web-based online courses are different from the classic e-Learning systems by the following characteristics: massive, open, online instructional course, certification based and enriched with learning activities. There are two forms distinct, the Canadian MOOCs came to be called cMOOCs (for "connectivist" ormany-to-many network) and the ones from Coursera, edX or Udacity came to be called "xMOOCs" (for "exponential" or one-to-many network). The specific application of Data Science (DS) in the education field is known as Educational Data Science (EDS) which works with data gathered from educational environments/settings to solve educational problems. There are a variety of technical methods to deal with big data, whose key is to synthetically make use of analysis techniques to provide learners for better study environment and learning advice; Such methods are: statistical analysis, Social Network Analysis (SNA), content analysis, data mining and data visualization. Next chapter are going to discuss the methodology which include, dataset description, proposed model and analysis techniques.

In this study, we used a quantitative design because numerical analysis nature of the data representation and analysis. The research started up by choosing the desired topic which is assessing student's performance antecedents and controllers of MOOC-based systems. Followed by defining the research problem and research objectives which achieved from digging inside the theoretical framework and literature review. For this study, the desired problem is investigating the determinants for student's performance by using a devoted model. Next is proposing the research model, desired hypothesis and building the survey based on the literature review and the chosen related theories. Followed by identifying and validating the data set used for analysis. Then the suitable data-set is analyzed for analysis of: demographic, descriptive and relations. Next, the findings were discussed in conjunction with the proposed hypothesis to achieve external validity of the proposed model. Finally, conclusions was discussed regarding: research objectives, research contributions and future recommendation. The data set used is the first of its kind jointly released by Harvard and MIT "Harvard XMITx Person-Course Dataset AY2013" in May, 2014-the first (de-identified) data available on MOOCs. The dataset consists of 641, 138 observations of 20 variables. Based on literature review and analysis of the dataset variables tweleve variables are chosen. For measurement purposes, two variables "grade" and "certification" which account for learning outcomes are available to the researcher. For explanatory purposes, five variables of students activity-based are chosen: interaction log interaction days, video interaction rate, chapter's interaction rate and forum interaction rate. For controlling purposed, four variables of learners trait-based are chosen: age, gender, country and education. Five hypothesis are examining direct relations while another four set of relations are examining the moderation affect.

REFERENCES

Abubakar, A.D., M.J. Bass and I. Allison, 2014. Cloud computing: Adoption issues for sub-Saharan African SMEs. Electron. J. Inf. Syst. Developing Countries, 62: 1-17.

Adler-Milstein, J. and A.K. Jha, 2013. Healthcares big data challenge. Am. J. Managed Care, 19: 537-538.

Aiden, E. and J.B. Michel, 2013. Uncharted: Big Data as a Lens on Human Culture. Penguin Publishing Group, London, UK., ISBN:9781101632116, Pages: 288.

Al-Fawaz, K., 2012. Investigating enterprise resource planning adoption and implementation in service sector organizations. Ph.D Thesis, Brunel Business School, Brunel University, England, UK.

Alla, M.M.S.O., 2013. The impact of system quality in E-learning system. J. Comput. Sci. Inf. Technol., 1: 14-23.

Arora, P. and F. Vermeylen, 2013. The end of the art connoisseur? Experts and knowledge production in the visual arts in the digital age. Inf., Commun. Soc., 16: 194-214.

Barak, M., A. Watted and H. Haick, 2016. Motivation to learn in massive open online courses: Examining aspects of language and social engagement. Comput. Educ., 94: 49-60.

Barocas, S. and H. Nissenbaum, 2014. Big datas end run around procedural privacy protections. Commun. ACM., 57: 31-33.

Boyd, D. and K. Crawford, 2012. Critical questions for big data: Provocations for a cultural technological and scholarly phenomenon. Inf. Commun. Soc., 15: 662-679.

Chang, R.M., R.J. Kauffman and Y. Kwon, 2014. Understanding the paradigm shift to computational social science in the presence of big data. Decis. Support Syst., 63: 67-80.

Chatti, M.A., A.L. Dyckhoff, U. Schroeder and H. Thus, 2012. A reference model for learning analytics. Intl. J. Technol. Enhanced Learn., 4: 318-331.

De Barba, P.G., G.E. Kennedy and M.D. Ainley, 2016. The role of students motivation and participation in predicting performance in a MOOC. J. Comput. Assisted Learn., 32: 218-231.