Learning and engagement assessment in MOOCs using multivariate methods and models ⋆

Maria Carannante[1], Cristina Davino[2], and Domenico Vistocco[3]

[1] Federica Weblearning Center for Innovation and Dissemination of Distance Education maria.carannante2@unina.it
[2] Department of Economics and Statistics cristina.davino@unina.it
[3] Department of Political Science domenico.vistocco@unina.it
University of Naples Federico II

Abstract. Massive Open Online Courses (MOOCs) phenomenon is the new frontier of online learning, where un-limited time and no location restrictions allow users to follow different strategies of learning. In the learning analytics literature there are many contributes dealing on how MOOC learners' behaviour affects their performance and influences reaching the course achievements. The present paper proposes a statistical model to analyse the relationship among learning, performance and engagement in a MOOC framework. As MOOCs offer different forms of learning, it is necessary to consider engagement and learning as multidimensional concepts, measurable using several indicators. The network of relationships is estimated through Partial Least Squares Path Modeling and differences in learners' behaviour according to age are also explored.

Keywords: Learning analytics · Multivariate modeling · Engagement.

1 Introduction

MOOCs phenomenon became popular in recent years as widely used learning tools in higher education institutes both in traditional and distance universities. Since MOOCs offer a variety of learning instruments, such as video lectures, multiple-choice quizzes, discussion forums and documents, the customization of learning analytics (LA) to the MOOC framework represents an important challenge [8]. One of the most cited definitions of LA is 'the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs' [11]. This paper aims to provide a contribution to the MOOC assessment exploring the effect of engagement and learning on students' performance. The methodological framework is represented by a consolidated multivariate method, Partial Least Squares Path modelling (PLSPM) [14,15], well-known in literature when it is necessary to measure a network of relationship among concepts not

⋆ Thanks to Federica Weblearning Center for Innovation and Dissemination of Distance Education, University of Naples Federico II, for providing the data

directly measurable. The analysis refers to one of the courses of IPSAMOOCs series offered by the Platform FedericaX, the EdX MOOCs platform of the "Federica WebLearning" Center at University of Naples Federico II [4], using data relating to socio-demographics characteristics and tracking log of 3,339 users' actions.

2 Model definition

Learning and engagement can be defined as two important drivers of students' performance. The present study aims at modeling the effect of these two components on the outcome of students attending a MOOC. As either learning and engagement and performance are complex concepts that cannot be directly measured by a single indicator, the first part of the study has been devoted to their conceptualisation and operativization [3]. The debate about **Learning** and **Engagement** concepts is still open, with many definitions deeply different each other and often overlapping. **Engagement** is a multidimensional concept and, in particular, we refer to the emotional engagement, defined as affective feelings for coursework, teachers or institutions [12]. **Learning** can be defined as the process of acquiring new, or modifying existing, knowledge, behaviours, skills, values, or preferences [1]. Conceptualisation, operativisation and measurement of learning and engagement required the definition of a proper set of indicators [3]. The **Learning** dimension has been structured into three sub-dimensions: *Frequency based activity*, *Time based activity* and *Interaction*. *Frequency-based data* are the simplest and most used data for synthesizing data from tracking logs [2]. Despite its simplicity, frequency can provide many useful information (e.g. the distribution of user events) to identify different behavioural patterns among learners. *Time-based activity* data give information about time spent studying. In literature, the quantity of time spent in learning activities is a predictor of students' performance. According to Hadwin et al. [5], analysing only the time spent in one or more learning activities is not enough, but it is necessary to analyse indicators about how a learner spent its time. The *interaction* sub-dimension relates to discussion forums activity and social learning activities. Discussion forum activity not only allows peer-to-peer learning and a direct interaction between teacher and learners, but it may help to reduce the drop-out rate. The **Engagement** dimension can be structured into two sub-dimensions: *Regularity* and *Procrastination*. The *Regularity* sub-dimension is related to the time-based activity dimension but from a different point of view as it measures how a learner spends his time on the platform and how he organizes his own learning roadmap.*Procrastination* is a key factor of MOOCs analysis; it can be viewed as the failure of the learner to organise its own learning process [7].

The paper explores the simplest structure relating the previous dimensions and sub-dimensions: the model is shown in Figure 1 where the direction of the arrows goes from the driver to the dependent concept (numbers on the arrows will be explained in Section 3). Further developments will regard the exploration of more complex structures with more connections (e.g. from engagement

to learning and vice-versa). For easiness of interpretation, the polarity of the indicators related to Procrastination has been reversed. The asterisk symbol has been assigned to the indicators with reversed polarity.

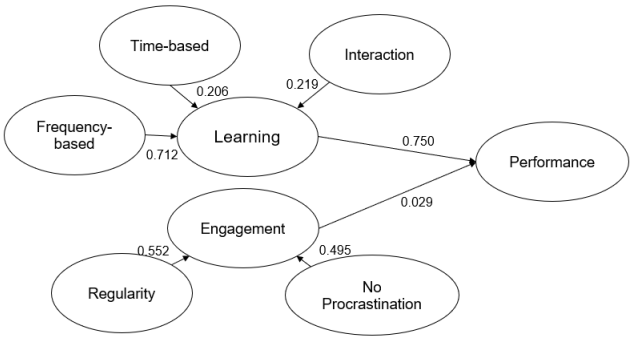

Fig. 1. Structural model for MOOCs learning assessment

In the model, **Performance** is the outcome dimension, that is it estimates the degree of effective learning of users. For this block, we consider a unique variable, the rate of correct responses with respect to total quizzes. The full list of the indicators adopted for each dimension and sub-dimension is shown in Table 1 (the last column will be described in Section 3, while numbers form 1 to 5 in the label column refer to the modules in which the course is structured).

3 Methodology and main results

PLS-PM is a consolidated statistical method able to model complex multivariate relationships among blocks of observed indicators (also known as manifest variables, MVs) and unobserved concepts named latent variables (LVs). A PLS-PM is made of a measurement (or outer) model relating each block of MVs to its corresponding LV and a structural (or inner) model connecting the LVs in accordance with a network of linear relationships. The PLS-PM algorithm allows to estimate separately the blocks of the measurement model and the structural model. In the PLS-PM terminology, **Engagement**, **Learning** and **Performance** and their related sub-dimensions represents LVs while the indicators described in Table 1 are the MVs. Figure 1 represents the adopted structural model. For further methodological details about PLS-PM see [6, 13, 14].

A preliminary analysis has been carried out to check unidimensionality and internal consistency of LVs taking into account the difference between the first and the second eigenvalue (the 1st eigenvalue is expected to be the only one greater than 1 and much higher than the second one), the Cronbach's α and the Dillon-Goldstein's ρ (grater than 0.7 in case of unidimensionality).

Table 1. Dimensions, indicators and coefficients

Dimension/*sub-dimen.*	Indicator	Description	Label	Weights
Performance	Rate of correct problems	correct problems \ problems	rate_correct	1
Learning				
Frequency-based activity	Rate of days active	days active \ course duration	active_days	0.178
	Average activities per day	activities \ course duration	ave_activities	0.233
	Rate of videos watched	different videos watched \ videos	rate_video	0.240
	Rate of rewatching	videos rewatched \ videos	rate_rewatched	0.231
	Rate of pages	different pages visited \ pages	rate_pages	0.224
	Average backward	backward on videos \ videos	rate_backward	0.126
Time-based activity	Average time spent on videos	time on video \ videos	video_time	0.319
	Average time spent a day	time active \ days active	ave_time	0.489
	Video completion rate	time on video \ video duration	video_completion	0.549
Interaction	Rate of forum pages	forums viewed \ forums	rate_forum	0.435
	Rate of participation to problems	problems tried \ problems	rate_problem	0.796
Engagement				
Regularity	Interval of days between activities	difference \ days active	interval_days	0.008
	Average activity time in a module	time \ modules	ave_time1	0.075
			ave_time2	0.199
			ave_time3	0.208
			ave_time4	0.203
			ave_time5	0.203
	Rate of return	previous modules activities \ activities	rate_return	0.097
	Lesson ordering	rate of videos seen in the right order	rate_ordering	0.253
	Difference in activity over time	skewness of activities	difference	0.121
No Procrastination	Time delay	day of first activity – realising day	delay1*	0.141
			delay2*	0.247
			delay3*	0.270
			delay4*	0.271
			delay5*	0.256

The estimation of the measurement part of the model allows to measure the importance of each indicator on the corresponding LV. Such measures, named outher weights, are shown in Table 1. All the coefficients are statistically significant at the 0.01 level using the classical t-test where standard errors are estimated through a bias-corrected bootstrap approach with 200 samples.

For *Frequency-based activity*, `active_days` and `rate_backward` give less contribution to the variable explication than the others, this means that learners pay more attention to the number and the kind of activities than to perform these activities as many days as possible, and they do not consider so important coming back on videos. For *Time-based activity*, the greatest contribution is given by `video_completion` and the lowest by `video_time`. This implies that it is not important how much time learners spend on videos, but if they watch the whole video. Participation to problems (`rate_problem`) has a very high impact on *Interaction* with respect to `rate_forum`. For *Regularity*, `interval_days`, `ave_time1` and `rate_return` give a very few contribution, this implies that learners' regularity is less conditioned by the interval of days between activities, by the activities of the first subsection and the return on previous subsections. Finally, for *No Procrastination* the size of the coefficients increases moving from the first module to the last, meaning that as the course progresses, it is much more important for students to be on time in starting an activity as soon as the learning materials are available.

The estimation of the structural part of the model allows to measure the impact of each sub-dimension on the related LV and of Engagement and Learning on Performance. Such weights, named path coefficients, are the numbers on the

arrows in Figure 1. The model suggests that **Performance** is mainly affected by **Learning** (coefficient equal to 0.75) while **Engagement** plays a quite irrelevant role (coefficient equal to 0.03) even not significant. On one hand to improve **Learning** it is advisable to act on actions related to *Frequency-based activity* which has the highest impact on Learning (coefficient equal to 0.71), almost three times greater than *Time-based activity* and *Interaction*. On the other hand, **Engagement** can be improved acting on *Regularity* (coefficient equal to 0.55) and reducing *Procrastination* (coefficient equal to 0.49) as well.

The analysis of the contribution (expressed by the path coefficient) of each determinant (i.e. each explanatory LV in the structural model) to the Performance can be deepened through an importance-performance map (IPMA) [9,10] (Figure 2) where the horizontal axis measures the so-called Importance, given by the path coefficients of the model, while Performance is the rescaled mean of each LV in the range 0–100 and measured on the vertical axis. The map can be divided into four quadrants, counterclockwise numbered from the one on the top right. Considering that the first quadrant is the area to keep, the second the overkill area, the third the low priority area and the fourth the critical area, it is evident that all the constructs based on frequencies can be improved and that the leverage two improve the **Performance** relies on the *Frequency-based activity* and more in general **Learning**.

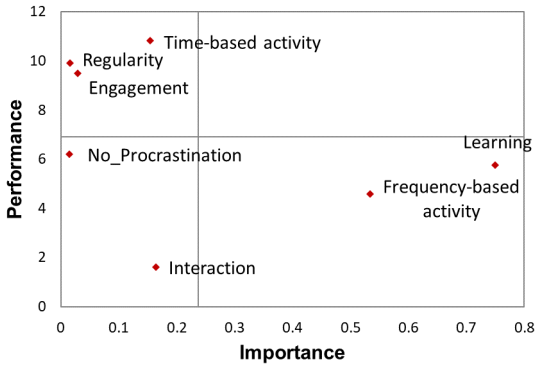

Fig. 2. Importance-Performance plot

A further exploration has been carried out to evaluate if the estimated relationships change in case of an observed heterogeneity, for example related to age. In order to identify differences by age we consider two groups of students: up to 24 years, that is the age of scholar or academic students, and over 24, that is the age of self-regulated learners. The group of over 24 years students significantly differs from the youngest students showing higher coefficients with respect to *Frequency-based activity*, *Interaction* and **Engagement**. The different role played by engagement is particularly interesting as it results a driver

of performance just for self-regulated learners. Further explorations of group differences can be realised using importance preference mapping.

In conclusion, we performed a multivariate model to estimate the concepts of Learning, Engagement and Performance and to measure the relationships among them. Moreover the proposed PLS-PM also allows to explore lower order relationships providing the impact of each sub-dimension and of each indicator.

References

1. Azevedo, R.: Defining and Measuring Engagement and Learning in Science: Conceptual, Theoretical,Methodological, and Analytical Issues. Educational Psychologist, **50**(1), pp. 84–94 (2015)
2. Beasley, R. E., Vila, J. A.: The identification of navigation patterns in a multimedia environment : a case study in an introductory course in artificial intelligence. Journal of Educational Multimedia and Hypermedia **1**, 209–222 (1992)
3. Carannante, M., Davino, C., Vistocco, D.: MOOCs learning assessment: conceptualisation, operativisation and measurement. In: Proceedings of INTED2019 Conference, pp 6694 – 6700. IATED academy, Valencia, Spain (2019)
4. EdX research guide, https://media.readthedocs.org/pdf/devdata/latest/devdata.pdf. Last accessed 1 Mar 2019
5. Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., Winne, P. H.: Examining trace data to explore self-regulated learning. Metacognition and Learning **2**(2), 107–124 (2007)
6. Henseler, J. Ringle, C. Sarstedt, M.: Using Partial Least Squares Path Modeling in International Advertising Research: Basic Concepts and Recent Issues. In: Editor, Okazaki, S. (eds.) Handbook of Research on International Advertising, pp. 252–276. Edward Elgar Pub (2012)
7. Howell A.J. and Watson D.C.: Procrastination: associations with achievement goal orientation and learning strategies. Personality and Individual Differences **43**(1), pp. 167–178 (2007)
8. Johnson, L., Becker, S. A., Estrada, V., Freeman, A.: Horizon Report: 2014 Higher Education.Austin TX, USA: The New Media Consortium, Austin TX, USA (2014)
9. Kristensen, K., Martensen, A., Grønholdt, L.: Customer satisfaction measurement at post Denmark: Results of application of the european customer satisfaction index methodology. Total Quality Management, **11**(7), 1007–1015 (2000)
10. Ringle, C.M., Sarstedt, M.: Gain more insight from your PLS-SEM results: The importance-performance map analysis. Industrial Management & Data Systems, **116**(9), pp.1865–1886 (2016)
11. Siemens, G., Long, P.: Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE review **46**(5), 30–41 (2011)
12. Sinatra, G.M., Heddy, B.C., Lombardi, D.: The Challenges of Defining and Measuring Student Engagement in Science. Educational Psychologist, **50**(1), pp. 1–13 (2015)
13. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., Lauro, C.: PLS path modeling. Computational Statistics & Data Analysis **48**(1), 159–205 (2005)
14. Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H.: Handbook of Partial Least Squares. 6th edition. Springer Berlin Heidelberg, (2010)
15. Wold, H.: Partial Least Squares. In: Editor, Kotz, S., Johnson, N. (eds.) Encyclopedia Statistical Science, pp. 1–13. Wiley, New York (1985)