# Analyzing Students' Behavior in UNED-COMA MOOCs

Llanos Tobarra[1], Salvador Ros[1], Roberto Hernández[1], Antonio Robles-Gómez[1], Rafael Pastor[1], Agustín C. Caminero[1], Jesús Cano[1], and Jordi Claramonte[2]

[1] Dpto. Sistemas de Comunicación y Control
ETSI Informática
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain
[2] UNED Abierta
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain

**Abstract.** This work presents an exploratory analysis about the students' behavior in the UNED-COMA MOOCs in order to detect and minimize drop-outs in MOOC courses. To achieve this, the following research questions are answered: a) what is the relationship between videos and students?; and b) what is the relationship between forums and students?. A set of indicators about students' behavior in MOOCs have also been extracted from this study, and a set of hypothesis have been analyzed. In particular, the amount and duration of videos do not affect the students' outcomes, according to the statistical analysis taken with our datasets, and possible drop-outs within the course. Some courses must include specific videos for a broader range of ages and country culture. In addition to this. the students' activity in courses does influence in drop-outs. As the period of the course advances, faculty should encourage students to participate in the interaction tools and visualize multimedia resources.

**Keywords:** Learning Analytics (LA); Exploratory Data Analysis (EDA); Drop-outs; Learning Indicators; Massive Open Online courses (MOOCs).

## 1 Introduction

Massive Open Online courses (MOOCs) have nowadays become an excellent platform to teach a great number of students at the same time with a distance methodology. As a consequence, a big amount of data is generated from students, interactions with the tools provided by MOOCs [2,10]. In particular, information from the multimedia resources and social interaction forums provided by the platform can be extracted and processed. Nevertheless, students can behave with different roles when they follow the course over time: *producers* and *consumers*. A producer is a student who interacts actively with the platform, but a consumer does not interact with the platform, she/he only uses the learning resources in an

isolated way [11]. Students with these roles would finish the course. In contrast, we have other kinds of students, who do not start the course or, even more important, they drop-out in some period the course [7]. For this reason, we are interested in detecting the reasons of this fact in order to establish mechanisms to minimize this negative impact and give some recommendations to faculty. Apart from this, this work aims at analyzing the students' learning in the context of the Learning Analytics (LA) topic [14].

On the other hand, the Spanish University for Distance Education, UNED, has started the UNED-COMA (Curso Online Masivo Abierto; in English, UNED-MOOC) [4] project for managing massive courses with multimedia videos and social interaction tools (like forums). In particular, UNED-COMA is an open initiative created in the 2012 year. The OpenMOOC [12] platform is employed for this purpose. The main objective of this initiative has been the exploration of the rich experience with the distance educational methodology employed at UNED. A very preliminar work focused on only languages courses in the OpenMOOC platform of UNED-COMA was presented in the LAK 2014 conference [13]. This work studies all datasets generated from 2012 to 2015 for all courses hosted in the OpenMOOC platform, by including statistical analysis of data and giving MOOCs general indicators only in the area of Foreign Language courses while we analyzed courses from all the different areas of UNED-COMA.

The OpenMOOC platform hosts a set of videos and discussion forums for each course, as main interactive available resources for students, being the principal interest of our study. Courses in OpenMOOC consist of units built by a set of knowledge pills. Pills are short videos linked to supplementary material (like documents, links, or exercises) and questions with their respective answers. Pills are classified in *normal*, *homework*, and *exams*. The difference among them is the time to study and the possibility or not to examine the answers before its deadline. Each course has associated an intelligent discussion forum, where students and faculty can discuss and collaborate on a unit/knowledge pill.

As for the tracking process, badges are automatically awarded to participants, recognising their contributions to the learning community. Other two kinds of accreditation are possible in OpenMOOC: 1) Credential: validation of having successfully finished and passed the course; and b) UNED-COMA certificate: a formal face-to-face exam in a associated centre of UNED.

With respect to the OpenMOOC structure, most of the courses are hosted as centralized MOOCs (xMOOCs). A small set of courses encourages self-learning and creating learning communities, sharing of intellectual work and open access to materials. We therefore need to consider the specific characteristics of MOOCs to see how quality can be described, assured and developed. For instance, a taxonomy of eight types of MOOC was proposed in [1], from a pedagogic point of view, not only the institutional perspective. Learning functionality, rather than MOOCs' origins, is exhaustively taken into account. Under this premise, OpenMOOC courses can be classified as *made MOOC* and *sync MOOCs* because of the creation of professional videos with the help of UNED infrastructure.

Apart from pedagogy, the courses in OpenMOOC can be classified in accordance with ten dimensions identified by Margaryan et al. [8] that give a best example of the nature of the course. See Table 1 for more details.

**Table 1.** OpenMOOC average course classification according to Margaryan et al [8].

| Dimension | Low | Medium | High |
|---|---|---|---|
| Problem-center | | x | x |
| Activation | | | x |
| Demonstration | | x | x |
| Application | | | x |
| Integration | | | x |
| Collective knowledge | | x | x |
| Collaboration | | x | x |
| Authentic resources | x | | |
| Feedback | | x | |

The rest of this work consists of four additional sections. First, a brief description of the methodology used in this work is given. Then, the most relevant details about the data pre-processing and contextualization are presented. Next section presents the research questions and hypothesis, and discusses the results obtained from the datasets. Finally, the primary conclusions and future work of this research are detailed.

## 2 Methodology

To carry out the analysis of the resulting research questions and hypothesis formulated in this work, the visual e-learning analytics process (VeLA) proposed by [5] has been adapted for our purposes. The VeLA model provides with a framework to process the data provided by OpenMOOC in order to prepare the information for visualization. From this process, Learning Analytics (LA) techniques can be employed in order to study the generated data.

The VeLA process taken here has been followed as detailed next. First of all, our dataset has been transformed in order to apply analytic models, particularly, an Exploratory Data Analysis (EDA), by including some statistical analysis. This initial dataset was composed by two different databases. On one hand, a PostgreSQL database contains all the static data related to OpenMOOC: information related to users, and courses structure and data platform. Additionally, a MySQL database has given support to the Q&A forums within our courses. Our OpenMOOC platform also uses a third database, a NoSQL MongoDB, to store activity data. At the time of this work, this database can not be accessed for its analysis.

In parallel with data analysis, we have mapped a portion of the data to visualize relevant patterns within the course resources, which could be relevant

for studying our research questions below. The generated results have provided feedback to post-processing the dataset and to repeat the process in order to discover new knowledge.

Next sections provided with details about the datasets, research questions and hypothesis, and the results obtained in this work.

## 3 Dataset description

Within the OpenMOOC platform, there are hosted 38 different courses and 99 course instances due to several repetitions of the courses during several years. The courses are classified in the following categories selected by the UNED-COMA administrators:

- General purpose (G).
- Languages (L).
- Science and Technology (ST).
- Social Service (SS).
- Laws (LW).
- Humanities (H).
- Economic and Business (EB).
- Psychology (P).

| Course cathegory | Male | Female | Unknown |
|---|---|---|---|
| Education, Economy and Bussiness | 0.92 | 1 | 0.17 |
| Education, Science and Technology | 2.26 | 3.74 | 0.54 |
| Economy and Bussiness, Social Service | 0.09 | 0.05 | 0.018 |
| Economy and Bussiness, Science and Technology | 5.42 | 4.19 | 0.82 |
| Economy and Bussiness | 3.36 | 5.17 | 0.82 |
| Education | 0.45 | 1.60 | 0.14 |
| General | 0.35 | 0.21 | 0.04 |
| Humanities | 0.61 | 1.48 | 0.18 |
| Laws, Economy and Bussiness | 0.62 | 1.14 | 0.19 |
| Languages | 20.07 | 31.426 | 4.06 |
| Psicology, Economy and Bussiness | 0.4 | 0.69 | 0.10 |
| Psicology, Social Service | 0.291338 | 0.533466 | 0.09 |
| Social Service, Economy and Bussiness | 0.09 | 0.132 | 0.02 |
| Social Service, Humanities | 0.869 | 1.299 | 0.199 |
| Science and Technology, Humanities | 0.637592 | 0.67 | 0.12 |
| Science and Technology, Laws | 0.0004 | 0 | 0 |
| Science and Technology | 1.70 | 0.66 | 0.23 |

**Table 2.** Percentage of students grouped by course categories during the period of 2012-2015.

Due to the complex nature of some of the courses, it has been very difficult to classify them into one specific category. Therefore, some of them are associated with two categories. The number of registered students is near 280.000.

In Table 2, the percentage of students per category group is represented. The OpenMOOC platform does not provide many personal information related to the students such as age or gender. In order to perform this analysis within the VeLA process, an automated classification to detect user gender from his/her name has been implemented. We can stablish that there is a strong interest in courses related to languages; which includes English, Spanish, and German. It is also remarkable the great female interest in courses of the OpenMOOC platform.

| year | new inscriptions | old students |
|---|---|---|
| 2012 | 11782 | 0 |
| 2013 | 158241 | 4193 |
| 2014 | 8343 | 14659 |
| 2015 | 1466 | 928 |

**Table 3.** Inscription and retention of students the period of 2012-2015.

If we pay attention towards the inscription of new students and the permanence of the students in the OpenMOOC platform represented at Table 3, there was an important interest in OpenMOOCs during 2013, but this initial enthusiasm dropped during 2014 and 2015 years. The lack of new students and the short permanence of the students in the platform are also two relevant issues that should be analysed carefully.

## 4 Research Questions and Discussion

In this section two research questions are detailed and discussed exhaustively. These ones focus on the students' patterns when visualizing videos hosted in the studied OpenMOOC courses, and their interaction within the social communication tools offered to them. In our particular case, the debate forums have been analyzed. A number of indicators belonging to OpenMOOC courses are also given along this section, in order to employ LA techniques when studying students' behaviors.

### 4.1 RQ1 - What is the relationship between videos and students?

As stated above, the most relevant elements within OpenMOOC courses are videos [6,9]. These resources are knowledge pills. Our OpenMOOCs do not offer students' tracking during the use of videos in the platform by students, but we are able to obtain some relevant information about videos within our OpenMOOC platform, since YouTube Analytic has been integrated in it. Videos hosted in the OpenMOOC platform are uploaded to YouTube by using principally two UNED-COMA accounts. Additionally, other small set of videos, less than the 10% of the platform videos, were uploaded with personal YouTube accounts of faculty. Therefore, we can not access to a full dataset of videos. Nevertheless, our

study is relevant enough to make conclusions, since a high percentage of videos have been analyzed.

Table 4 shows the geographical distribution, in terms of age and sex, of visitors for the videos hosted in the principal video channels employed by the OpenMOOC platform. This information can only be gathered when the user visualizes a video, while logged with a Gmail account. This is not a problem, since almost 70.000 students use it to access the OpenMOOC platform, according to the data stored in the database. They represent

| Country | Age | Male | Female |
|---|---|---|---|
| Colombia | 65- | 0.0 | 0.0 |
| | 55-64 | 0.0 | 0.0 |
| | 45-54 | 25.0 | 0.0 |
| | 35-44 | 7.1 | 0.0 |
| | 25-34 | 32.1 | 0.0 |
| | 18-24 | 28.6 | 0.0 |
| | 13-17 | 3.6 | 3.6 |
| Spain | 65- | 1.8 | 0.0 |
| | 55-64 | 3.3 | 0.0 |
| | 45-54 | 37.0 | 5.1 |
| | 35-44 | 12.0 | 4.0 |
| | 25-34 | 20.3 | 9.1 |
| | 18-24 | 3.6 | 4.0 |
| | 13-17 | 0.0 | 0.0 |
| Mexico | 65- | 0.0 | 0.0 |
| | 55-64 | 18.2 | 0.0 |
| | 45-54 | 0.0 | 0.0 |
| | 35-44 | 24.2 | 0.0 |
| | 25-34 | 45.5 | 0.0 |
| | 18-24 | 9.1 | 0.0 |
| | 13-17 | 3.0 | 0.0 |
| Peru | 65- | 0.0 | 0.0 |
| | 55-64 | 0.0 | 0.0 |
| | 45-54 | 2.6 | 0.0 |
| | 35-44 | 0.0 | 0.0 |
| | 25-34 | 10.3 | 0.0 |
| | 18-24 | 53.8 | 33.3 |
| | 13-17 | 0.0 | 0.0 |

**Table 4.** Geographical distribution of video visitors by age and sex.

As observed in Table 4, the percentage of visits are shown for each origin country of videos, and divided by the age in ranges and sex (male or female) of student visitors. On one hand, we would like to highlight the great impact of our courses has in American countries of Spanish tongue. Age range is also important in order to detect the most convenient public of our courses, and to study how to improve courses to attract them to the rest of visitors. For instance, visitors

between 25 and 34 years old are the most interested in our courses in Peru. As for the particular sex of visitors, although the volume of students enrolled in courses is higher for women, men are more common as video visitors. This can become a contradiction, it is probably that women watch videos without a Gmail account session.

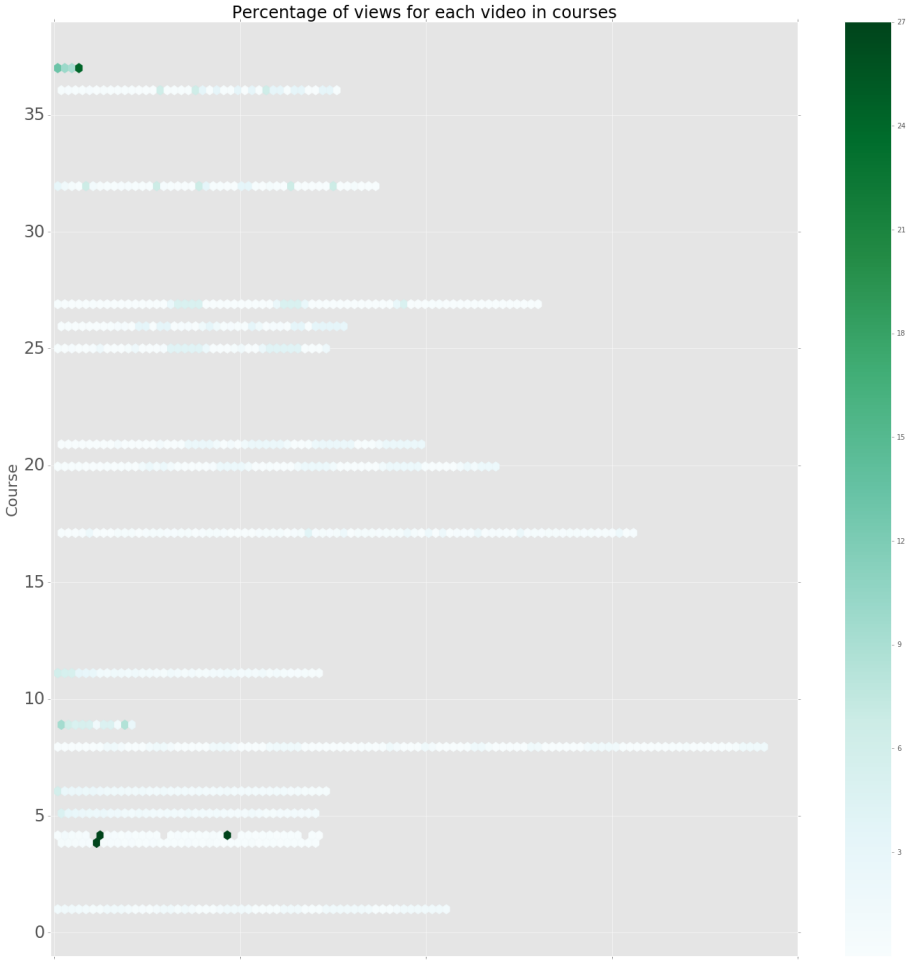


**Fig. 1.** Course activity in terms of number of video visits.

Figure 1 shows the course activity in terms of number of video visits with the same order scheduled in the OpenMOOC courses. From the 38 courses hosted in the OpenMOOC platform, each course contains a particular number of videos, and it can differ among courses (in this case, from 7 videos the shorter one to 191 videos the longer one). The number of visits for each video in the courses

have been normalized to make possible a comparative study. The colour of each point in Figure 1 is proportional to the number of video visits. A dark green color indicates a big amount of visits, whereas a clearer point indicated a lower number of visits. In our particular case, in most of the courses we can find an predefined pattern, which consists on a series of videos with a few visits followed by a videos very visited. This particular video usually contains guidelines amount the course assessment, so it gets more visits.

Taking into account the previous information several hypothesis have been formulated:

- *Hypothesis 1:* shorter videos are better for the students' outcomes.
- *Hypothesis 2:* courses with fewer videos are more successful that those ones with a great number of videos.

We are going to study these hypothesis by means of a set of data variables or indicators. Now, a set of candidate variables (indicators) related to the success or failure of students in the OpenMOOC platform are given:

- *Number of videos per course (*$|videos|$*)*: it is the amount of videos that each student has to watch to gather the minimum required knowledge.
- *Mean duration of videos per course (*$\overline{duration}$ *)*: it is the mean duration in seconds of available videos in the course.
- *Mean amount of visits per video in the course (*$\overline{viewcount}$ *)*: Each time a student watches a video, this YouTube counter increases in one unit.
- *Mean amount of likes per video in the course (*$\overline{likecount}$ *)*; Each time a student likes a video, this YouTube counter increases in one unit.

**Table 5.** Candidate variables (indicators) for video resources.

| Variable | Mean | $\sigma$ | Max. | Min. | $\rho$ | p-value |
|---|---|---|---|---|---|---|
| $\|videos\|$ | 81,08 | 45,82 | 229 | 7 | -0,012 | 0,91 |
| $\overline{viewcount}$ | 37568,79 | 39838,23 | 146970,51 | 354,87 | 0,094 | 0,40 |
| $\overline{likecount}$ | 30,40 | 114,13 | 601,91 | 0,23 | -0,005 | 0,64 |
| $\overline{duration}$ | 163,55 | 107,68 | 559,23 | 4 | 0,01 | 0,91 |

All the candidate variables obtained in this video category are not linked to students' behavior. This evidence is supported by statistical analysis when we perform the Spearman's rank correlation test [3] with the number of students that successfully finished the course. See Table 5 to observe the particular results of the candidate variables. It is clear that the *p-value* of the variables is greater than 0,001 and the rank correlation among variables is very small. Thus, we can conclude these ones are not good indicators to associate the visualization of videos and passing the course. Although it seems logical to think that the number of videos or its duration have an impact in the learning outcome of the students, according to our data, they are not significant enough to correlate them.

### 4.2 RQ2 - What is the relationship between forums and students?

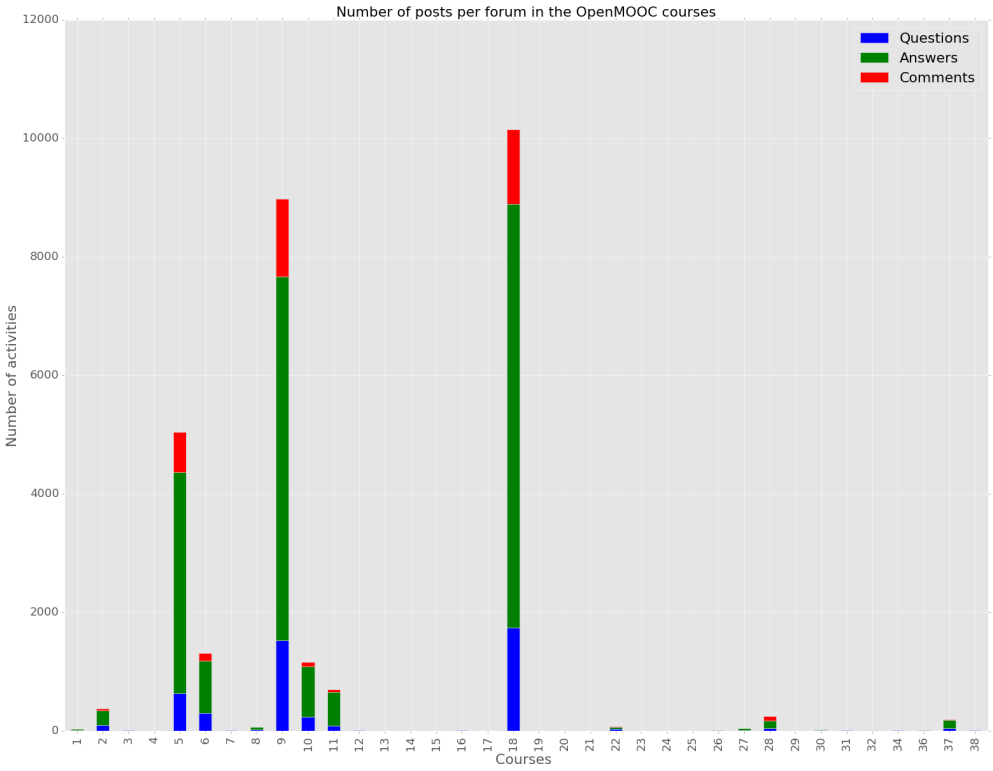


**Fig. 2.** Types of activities within forums of each course.

The debate forums in the OpenMOOC platform are represented as Q&A, in a similar manner as StackOverflow platform. In this sense, students can start questions (*question*), answer to existing questions (*answer*), or enforce some previous answer (*comment*).Figure 2 represents the activity volume in forums associated to each OpenMOOC course. From the analyzed data, the most popular Q&A activity has been *answer*, and there are a few comments. We would also like to highlight that some courses have not had any activity in it. This fact is very significant and, faculty should encourage his/her students to participate in the platform in order to enrich the learning process. The most active courses in forums are also the ones related to languages with a high difference of messages.

Figure 3 shows the daily activities when the debate forums are analyzed per course. The periods of more activity is when a course starts (they start at different periods of time). In particular, we can observe that December and March are the months with a higher activity, the dates coincides with the starting of language courses.

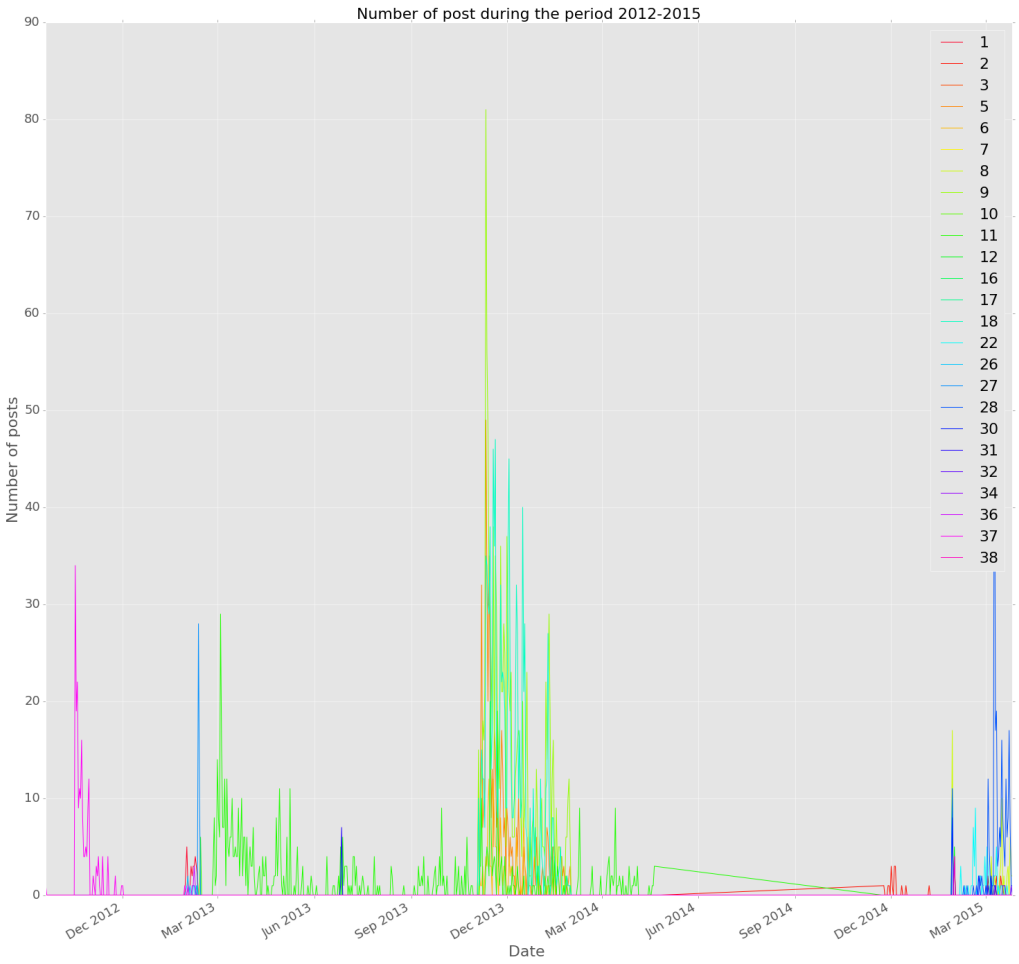


**Fig. 3.** Daily activity of debate forums for the OpenMOOC courses.

According to debate forums, the candidate indicators to become relevant indicators of students' behavior are the following ones:

- *Number of activities (|posts|)*: it is the number of total post messages for each course.
- *Number of questions (|questions|)*: it is the number of started posts for each course.
- *Number of answers (|answers|)*: it is the number of answered posts for each course.
- *Number of comments (|comments|)*: it is the number of comments for each course.
- *Percentage of the number of students (pauthors)*: it is the percentage of the total of the number of students of the course that takes part of the forum's activities.
- *Average number of activities per user ($\overline{activitiesuser}$)*: it is the mean activity of users for each course.
- *Average number of activities per day ($\overline{postperday}$)*: it is the mean number of posts per day for each course.

The main hypothesis that we can formulate over forum data is that the activity is directly related to the success of the course. We can formulate as follows:

- *Hypothesis 3:* a elevated number of activities in the forums of the course is related to the student success within the course.

According to the evidence obtained by means of the Spearman's rank correlation test [3] (see Table 6), the most correlated indicators to the badge awarded are the number of activities in the forum per course, the number of answers provided per course, and the number of posts per day in each course. These indicators have been analyzed for those courses where there were activity. Higher values of these variables are proportional related the number of students that obtained a badge. Therefore, these indicators should be followed in order to detect early drop-out of students, and to encourge students with additional questions, resources, and so on.

We focus our attention towards the more active students, we can not correlate the success of the students that not take an active part in the forums. Although there are works that have been analyze this type of students [10,11].

## 5 Conclusions

The main objective of this work has been looking for the reasons why a big amount of students drop-outs in MOOCs courses and studying the information offered by these types of courses. The UNED-COMA courses have been chosen for this study. In this sense, the VeLA methodology has been adapted for our purposes in order to prepare the data to make an exploratory data analysis,

**Table 6.** Candidate variables (indicators) for data in debate forums.

| Variable | Mean | $\sigma$ | Max. | Min. | $\rho$ | p-value |
|---|---|---|---|---|---|---|
| $\|posts\|$ | 1226,82 | 2845,02 | 10144 | 1 | 0,66 | 0,0006 |
| $\|questions\|$ | 207,43 | 474,42 | 1744 | 1 | 0,48 | 0,021 |
| $\|answers\|$ | 862,26 | 1994,44 | 7150 | 0 | 0,70 | 0,0001 |
| $\|comments\|$ | 157,13 | 381,93 | 1316 | 0 | 0,58 | 0,004 |
| $pauthors$ | 8,49 | 10,83 | 32,89 | 0,03 | 0,29 | 0,19 |
| $\overline{activitiesuser}$ | 3,28 | 1,55 | 6,05 | 1 | 0,366 | 0,086 |
| $\overline{postperday}$ | 0,27 | 0,52 | 1,83 | 0,001 | 0,63 | 0,0013 |

which includes analyze the data statistically, so extracting new knowledge from our MOOC courses.

The lack of one of the databases from the platform, that contains data related to the student's activities, has been a great impact in our results. This database could allow us to correlate the results of the partial evaluations with the performance of the students inside the course and the videos. So, the conclusions of this analysis could be improved.

In our case, the OpenMOOC platform has been employed, which focuses on multimedia resources and interaction tools provided to students during their period of learning. Two main questions have been analyzed in this work: a) what is the relationship between videos and students?; and b) what is the relationship between forums and students?. Many courses must include additional videos for a broader range of ages, and thinking the destination country culture. As the period of the course advances, faculty should also encourage students to participate in debate forums, use the multimedia resources; that is, maintaining the student alive in the course.

A set of hypothesis have also been proposed and discussed from the detected candidate variables (indicators) from datasets. This indicators are about students' behavior in MOOCs given both from the point of view of videos and debate forums. For instance, we have concluded in a statistical way that the duration and amount of videos in an OpenMOOC platform do not affect the students' outcomes (that is, possible drop-outs in the course), according our datasets, although their activity does. This way, these indicators could be useful as basis for further studies in any MOOCs platforms.

## Acknowledgments

The authors are specially grateful to UNED-COMA for facilitating the access to the needed data for this research study. Also, authors would like to acknowledge the support of the European research project ERC-2015-STG-679528 POST-DATA, and the local project (2014I/PPRO/031) from UNED and Banco Santander; and the Region of Madrid for the support of E-Madrid Network of Excellence (S2013-ICE2715). The authors also acknowledge the support of SNOLA,

officially recognized Thematic Network of Excellence (TIN2015-71669-REDT) by the Spanish Ministry of Economy and Competitiveness.

## References

1. Clark, D.: Moocs: a taxonomy of 8 types of moocs. avaible at http://donaldclarkplanb.blogspot.se/2013/04/moocs-taxonomy-of-8-types-of-mooc.html (2013), last access october 2016
2. Clow, D.: Moocs and the funnel of participation. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 185–189. LAK '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2460296.2460332
3. Coefficient, S.R.C.: The Concise Encyclopedia of Statistics, pp. 502–505. Springer New York, New York, NY (2008), http://dx.doi.org/10.1007/978-0-387-32833-1_379
4. COMA, U.: https://coma.uned.es/ (April 2017), last access: april 2017
5. Conde, M.Á., García-Peñalvo, F.J., Aguilar, D.A.G., Therón, R.: Exploring software engineering subjects by using visual learning analytics techniques. IEEE-RITA 10(4), 242–252 (2015), http://dx.doi.org/10.1109/RITA.2015.2486378
6. Giannakos, M.N., Chorianopoulos, K., Chrisochoides, N.: Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. The International Review of Research in Open and Distributed Learning 16(1) (2015), http://www.irrodl.org/index.php/irrodl/article/view/1976
7. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 170–179. LAK '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2460296.2460330
8. Margaryan, A., Bianco, M., Littlejohn, A.: Instructional quality of massive open online courses (moocs). Computers & Education 80, 77–83 (January 2015), http://oro.open.ac.uk/46374/
9. Muñoz Merino, P.J., Ruipérez-Valiente, J.A., Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado Kloos, C.: Precise effectiveness strategy for analyzing the effectiveness of students with educational resources and activities in moocs. Comput. Hum. Behav. 47(C), 108–118 (Jun 2015), http://dx.doi.org/10.1016/j.chb.2014.10.003
10. Mustafaraj, E., Bu, J.: The visible and invisible in a mooc discussion forum. In: Proceedings of the Second (2015) ACM Conference on Learning @ Scale. pp. 351–354. L@S '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2724660.2728691
11. Nagel, L., A.S.Blignaut, Cronjé, J.: Read-only participants: A case for studetn communication in online classes. Interactive Learning Enviroments 17(1), 37–51 (2009)
12. OpenMooc: http://openmooc.org/ (June 2013), last access:april 2017
13. Santos, J.L., Klerkx, J., Duval, E., Gago, D., Rodríguez, L.: Success, activity and drop-outs in moocs an exploratory study on the uned coma courses. In: Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. pp. 98–102. LAK '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2567574.2567627

14. Siemens, G., Baker, R.S.J.d.: Learning analytics and educational data mining: Towards communication and collaboration. In: Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge. pp. 252–254. LAK '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2330601.2330661c