A learning analytics methodology for understanding social interactions in MOOCs

Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J. Muñoz-Merino, *Senior Member, IEEE,* Iria Estévez-Ayres, and Carlos Delgado Kloos, *Senior Member, IEEE*

Abstract—One of the characteristics of MOOCs (Massive Open Online Courses) is that the overall number of social interactions tend to be higher than in traditional courses, hindering the analysis of social learning. Learners typically ask or answer questions using the MOOC forum. This makes messages a rich source of information, which can be used to infer learners' behaviour and outcomes. It is not feasible for teachers to process all forum messages and automated tools and analysis are required. Although there are some tools for analysing learners' interactions, there is a need for methodologies and integrated tools that help to interpret the learning process based on social interactions in the forum. This work presents the 3S (Social, Sentiments, Skills) learning analytics methodology for analysing forum interactions in MOOCs. This methodology considers a temporal analysis combining the social, sentiments and skill dimensions that can be extracted from forum data. We also introduce LATᴈS, a Learning Analytics tool for edX / Open edX related to the three dimensions (3S), which includes visualisations related to the three dimensions to guide the proposed methodology. We apply the 3S methodology and the tool to a MOOC on Java programming. Results showed, among others, the action-reaction effect produced when learners increase their participation after instructor's events. Moreover, a decrease of positive sentiments over time and before deadlines of open-ended assignments was also observed and that there were certain skills which caused more troubles (e.g., arrays and loops). These results acknowledge the importance of using a learning analytics methodology to detect problems in MOOCs.

Index Terms—Discussion forums, Distance Learning, Learning Environments, Visualization

——————————— ◆ ———————————

1 INTRODUCTION

IN the last five years, education has experienced a significant revolution with the appearance of MOOCs (Massive Open Online Courses). MOOCs open the possibilities for learning about almost anything with formats which typically include video lectures, activities, reading materials, forums, etc. [1]. However, unlike in traditional courses, MOOCs do not include face-to-face interactions with teachers or classmates. Instead, there are several ways of communication, which include forums, social networks, or blogs, among others [2]. Particularly, the forum of the MOOC is the preferred source [2] of communication and where most interactions between learners occur [3]. Because of its relevance, all MOOC platforms include a built-in forum.

In MOOC forums, learners are invited to ask questions and answer those posted by their coursemates. Learners can also share interesting insights about the course topics, and can even provide their impressions and affective states throughout the course. Teachers can also participate to support learners. Nevertheless, the large number of enrolees in MOOCs implies that there can be hundreds of messages every day and thousands overall, and it is

not feasible for teachers to read or reply to all posts [4]. Because of that, it is interesting to automate the management of messages, so that instructors can get meaningful information related to the forum activity easily. One possible benefit of this automatisation is that teachers will be able get information regarding how the forum works, which can serve to identify what parts of the course cause more trouble to learners, so these parts can be re-designed, including e.g., additional supporting materials. Therefore, this automatisation can produce a twofold improvement at both teaching and learning levels: teachers can improve their methodologies to adapt courses to learners' needs, and learners can benefit from improved re-designed courses.

One key aspect is that the starting point for the analysis of social interactions is usually raw data about thousands of messages posted in the MOOC forum, and it is important to process them automatically and design meaningful visualisations that properly communicate useful information about the learning context to the stakeholders. One possible way to provide these visualisations is though dashboards, which allow presenting the information in single views to be seen at a glance [5]. Dashboards have already been widely used in education (e.g., [6]) and, particularly, MOOC platforms typically use them to offer visualisations about learners' profiles and their interactions with learning resources. For example, edX offers edX Insights [7], which provides graphics mainly related to demographic information, learners' engagement and performance. Similarly, Coursera has a dashboard with live view of the data, including infor-

————————————————

- *P.M. Moreno-Marcos is with the Universidad Carlos III de Madrid, Leganés, ES 28911. E-mail: pemoreno@it.uc3m.es*
- *C. Alario-Hoyos is with the Universidad Carlos III de Madrid, Leganés, ES 28911. E-mail: calario@it.uc3m.es*
- *P.J. Muñoz-Merino is with the Universidad Carlos III de Madrid, Leganés, ES 28911. E-mail: pedmume@it.uc3m.es*
- *I. Estévez-Ayres is with the Universidad Carlos III de Madrid, Leganés, ES 28911. E-mail: ayres@it.uc3m.es*
- *C. Delgado Kloos is with the Universidad Carlos III de Madrid, Leganés, ES 28911. E-mail: cdk@it.uc3m.es*

mation about engagement or reach (e.g., number of learners committed to complete, audit, etc.), although the granularity of the information is sometimes limited [8]. However, forum activity is often neglected and, in most cases, only simple data about the number of posts is presented. In addition, there is a need for methodologies that serve as a reference on how to combine, analyse and interpret the different forum indicators and visualisations. These methodologies can be used to extract useful conclusions about the learning process, with the aim to detect the problems and make proper decisions to improve the course. In this direction, the objectives of this work are the following:

1) Propose a new methodology, named 3S (Social, Sentiments, Skills) for analysing, interpreting and reaching conclusions about the learning process in MOOCs based on forum activity.

2) Create a specific visual analytics tool (named LATƎS – Learning Analytics Tool for the 3S methodology) for the analysis of forum activity according to the proposed methodology. The tool should be compatible with edX / Open edX since this is the most extended open source MOOC platform.

3) Analyse a real specific MOOC with the 3S methodology and LATƎS. This course shall be representative, containing sufficient enrolees and forum posts.

With these objectives, this article is focused on the methodology, while the tool serves to apply the methodology in a real scenario. The tool is a specific instance to support the proposed methodology although there might be other possible tools that address this methodology. The structure of the paper is defined as follows. Section 2 provides a background on researchs about visualisation systems, social learning and, particularly, on contributions with deal with forum-related data. Section 3 describes the proposed 3S methodology. Details about LATƎS are explained in Section 4, while the case study is presented in Section 5. Finally, the main conclusions and future research directions are provided in Section 6.

2 RELATED WORK

Our 3S proposal relies on the use of dashboards for forum interactions. In this section, the most important articles related to visualisations systems in education are covered. First, a general overview is presented. Next, articles that deal with data taken from course forums, especially those which provide visualisations on this data, are discussed.

2.1 Visualisation systems in Education

Visualisation systems are designed to provide information obtained from the data to the stakeholders in a visual way. Verbert et al. [9] presented a model with four elements to define dashboards: awareness, self-reflection, sense making, and impact. Dashboards are deployed to support teachers to obtain an overview of what is happening in the course (awareness) to reflect themselves on how their educational practices (self-reflection), so that they can identify learners at risk (sense making) and perform interventions to reverse the situation (impact).

However, teachers are not the only target users of dashboards. Schwendimann et al. [10] performed a systematic review of the state of the art and provided a classification of the main stakeholders. They found that teachers are the primary users of dashboards, followed by students, administrators and finally researchers. Their analysis also focused on the scenarios where data for visualisations were taken. In this case, possible scenarios were formal, non-formal and informal learning [11]; formal learning was reported to be the predominant with 91% of the analysed articles referring to this type of scenario. The analysis also showed that the most common visualisations were bar charts, followed by line charts, tables and pie charts. Moreover, indicators used to create the visualisations were classified in six categories: (1) learner-related to describe the learner, (2) action-related to present information about the actions performed by the learner, (3) content-related to give information about the materials users interacted with, (4) result-related to show details about the learning outcomes, (5) context-related to give a picture about the learning context, and (6) social-related, to provide information about how learners interact with each other. Among these categories, the latter will be especially relevant in this paper.

Despite data is very important, the purpose of the dashboard also becomes relevant. Verbert et al. [12] classified dashboards depending on if support face-to-face classes, face-to-face group work, or online or blended learning. For example, Yu et al. [13] presented a dashboard showing the attitudes detected when students nod or shake their heads (face-to-face classes); Do-Lenh [14] proposed TinkerBoard to visualise groups activity when e.g., doing questionnaires or launching simulations (face-to-face group work); and Ali et al. [15] described the tool LOCO-Analyst, aimed at providing educators with feedback on the relevant aspects of the learning process taking place in an online platform (online learning). In online platforms, Ruipérez-Valiente et al. [16] also developed ALAS-KA to provide visualisations about learners (aggregated and individualised) in the Khan Academy. However, Verbert et al. [12] recognised the lack of evaluation of dashboards and, in terms of sources, they observed that dashboards rely mainly on systems logs.

Another key aspect of a dashboard is the presented information. There can be many types of contents to be displayed, such as demographic information, interactions with exercises, videos, forums, and so on. For instance, Shi et al. [17] developed VisMOOC, a dashboard which took video-related data from Coursera and showed visualisations, such as the parts of the video which had more reproductions, pauses, etc. Apart from that, edX has its own tool called edX Insights [7], which provides information related to users' activity (interactions with the platform in general, interactions with videos, results in assignments and engagement), and demographic information (age, level of education, gender, geographic information or data related to the number of enrolled users). Data are presented both aggregated and for individual learners. However, the information related to forum activity in both Coursera and edX is very limited. For example, edX only offers a graphic which indicates the

number of messages over time, without the possibility of filtering by date. Another tool used for visualisation with Open edX data is ANALYSE [18] . This tool provides 12 visualisations related to the interaction of learners with different resources, including videos, exercises or general course activity. This tool allows, for example, analysing what specific seconds of the video have been watched more times. However, ANALYSE also does not provide visualisations related to forum activity.

In this work, we propose a methodology based on the analysis of visualisations to indicate how instructors should analyse these forum activities. In addition, a tool is created according to this methodology. We propose a tool, which is focused on edX / Open edX, as it is an open source platform with a wide adoption in MOOCs. Moreover, the tool will be a technological innovation because it will cover the aforementioned limitation regarding the analysis of forum interactions in dashboards, and particularly in edX. Nevertheless, it is important to note that the main contribution of this article is the methodology, which defines the steps to interpret statistical analysis. Finally, the analysis of MOOCs with the tool is also a research opportunity according to the state of the art since the conclusions of the work by Schwendimann et al. [10] indicated the lack of research in non-formal settings.

2.2 Social Learning Analytics

Social learning is a key element in MOOCs. According to Shum and Ferguson [19], learners use social learning to clarify their intentions, ground their learning and engage in learning conversations. All these purposes share the medium of communication, which is typically the forum of the course, although sometimes social networks, blogs and other tools are used as well. Because of that, forum activity can provide a lot of meaningful information regarding learners' behaviours and there have been several contributions which analysed forum data in MOOCs. For instance, Alario-Hoyos et al. [20] considered data from different social tools that can be used in MOOCs, such as the forum, Facebook or Twitter, among others. They concluded that the forum was the preferred tool for MOOCs.

Another example was the work by Chen et al. [21], which focused on identifying the personality of learners in the following categories: openness, extraversion, conscientiousness, agreeableness and neuroticism. Moreover, Yang et al. [22] developed a model to identify confusion of learners (e.g., when a student says, "I'm stuck"). In addition, there have been contributions which have focused on predicting if there will be intervention from an instructor in a forum message or not [23], or have classified messages according to different categories, such as the course aspect [24] (i.e., if the message is about a video, a quiz, how to obtain the certificate, etc.) or the sentiments learners show in the messages [25] (commonly known as sentiment analysis). More related with this contribution, authors previously used forum-related variables related with social and sentiment aspects to predict grades although results showed they were not good predictors [26].

The abovementioned contributions focus on obtaining

conclusions from forum-related interactions, but the way to transmit this information to the stakeholders is also important. Because of that, there are also contributions in education which provided visualisations regarding forum activity. For example, authors in [27] developed ConVis, a visual text analytic system for blogs which included a visualisation relating authors, main topics of a thread and sentiments of the posts. However, that tool had difficulties to scale with massive data, such as in MOOCs [28]. Moreover, the SNAPP tool allows users visualising the network of interactions from forum posts and replies [29].

In MOOCs, Chitsaz et al. [30] designed a dashboard focused on FutureLearn, which showed the number of comments separated by their context (i.e., if messages were related to articles, videos, etc.) and included a graph on sentiment analysis. Moreover, edX allows visualising the timeline of forum posts and also includes functionalities for sorting and displaying messages by number of votes, number of responses and date. Nevertheless, this information has two main drawbacks: no information is provided on the most relevant topics or resources which may contain errors or cause trouble to learners; and no date filtering is enabled, which makes messages with more responses are always at the beginning of the list, although they might be older and probably solved.

The previous examples provide ways to visualise or display information about forum data, but this information is usually about some specific indicators and there are not methodologies about how to combine these data to take advantage. The information can only improve educational practises if instructors really understand and use it. There is a need of methodologies to analyse social interactions in MOOCs integrating different sources. To that end, this work aims to present a novel methodology to guide teachers when analysing forum-related data in MOOCs. Currently, apart from dashboards, there are many articles regarding the analysis of forum data, but an integral solution is needed to combine different approaches and raise conclusions about the learning processes and the problems learners face in this kind of courses. The methodology presented is aimed to do that and provide new solutions to support instructors in their MOOCs.

3 METHODOLOGY

This section presents the 3S (Social, Sentiments, Skills) data-driven methodology to analyse forum interactions in MOOCs. The aim of this methodology is to provide a guide for instructors to analyse the forum data of their courses, following a systematic sequence of steps. Although other sources (e.g., videos, exercises) can be subject of the analysis, this methodology only focuses on forum interactions because of the lack of research works using forum activity, unlike with other data sources. In addition, the communication between learners is very intensive in c-MOOCs and forum activity becomes very relevant. In this section, the dimensions in which learners' messages are structured are presented first. Next, the description of the methodology is detailed, including the steps to follow in order to consider all the dimensions.

3.1 Forum interaction dimensions

Forum posts in a MOOC can provide a lot of heterogeneous information about the learner. For example, the number of times learners interact can be analysed, or the evolution of their participation (as it can be done with other variables like those extracted from videos, exercises, etc.). However, the semantic analysis of the text of the messages can also provide insightful information regarding which contents of the course cause more difficulties to the learner or what learners' mood is throughout the course. As there are different aspects to consider in forum data, the 3S methodology structures the information in three dimensions to facilitate the analysis:

- **Social**. This dimension is related to the interactions of the learner with others in the forum (social activity). Some of the aspects which can be considered in this dimension (like those e.g., in [31]) include: the number of messages posted, the number of contributors of these messages, the number of responses and votes posts have, and the response time between the first message in the threads and their first response.
- **Sentiments**. The second dimension aims to understand learners' mood. Learners' sentiments can affect the learning processes [32] and text mining techniques can be used to determine the polarity of messages (e.g., positive/neutral/negative).
- **Skills**. The third dimension is related to the skills of the course that learners mention in their posts. Skills which appear more frequently in the forum are more likely to pose difficulties to learners as one of the main uses of the forum is asking questions [33], (e.g., when the students are in trouble in the tasks [34]).

These 3S dimensions are the basis for the analysis of forum data, while there is a fourth element that must be considered: time. If the whole period of the course is considered, an instructor could have an idea of how many messages are, what skills produce a higher activity, etc., but the instructor would fail to observe if there are specific periods where negativity suddenly increases, where the activity rises or drops considerably, or where certain skills lead to more discussion than others. This information can help for the re-design of specific parts of the course where the outcomes are not as expected. Figure 1 depicts a schema of the dimensions of the 3S methodology, including time; the relationship between dimensions is represented with arrows. This means that the methodology is intended to be a sequential process, so instructors must start with the social dimension, and follow with sentiments and skills dimensions to cover them all.
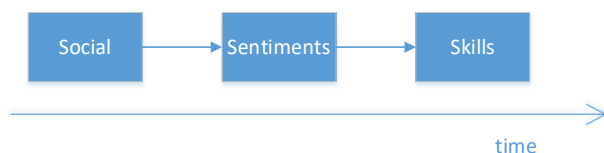

Fig. 1. Overview of the 3S methodology flow

3.2 3S Learning Analytics Methodology

The 3S learning analytics methodology has been designed to support instructors in the analysis of forum data of their MOOCs. The methodology is intended to be gener-alisable to any kinds of MOOCs, although there are some variations depending on if the course is delivered in synchronous (instructor-paced) or asynchronous (self-paced/learner-paced) modes. The reason is that in an asynchronous course, it is not possible to map contents to specific periods of time (e.g., concept X should be learnt in calendar week Y) as learners can advance in the course at different paces and even enrol in distant dates, unlike in synchronous courses where all learners are supposed to do the same activities in the same calendar weeks. Nevertheless, note that the interpretation of the methodology is subjective regardless the kind of MOOC because there can be many factors (e.g., course design and pedagogy) that can affect the interpretation, and each teacher should interpret the results based on the context of his course.

3.2.1 Step 1: Selection of relevant intervals of time

The first step of the 3S methodology consists on selecting the intervals of time where there has been a relevant trend change in forum activity (e.g., number of messages per day) for a deeper analysis. To do this task, it is necessary to obtain e.g. a graphic with the evolution of number of posts over time. When analysing this graphic, an instructor should be aware that forum activity usually drops after the first days of the course (e.g., see [2]). Because of that, if a significant decrease in the number of contributions is detected after the first or two first weeks, then that should not necessarily imply the interval needs a deeper analysis, as it can be justified by the fact that many learners enrol just to have an overall look at the content.

After the beginning of the course, if any patterns where forum activity suddenly increases or drops are identified, then they should be detected. In this case, the instructor should take the interval of time where each pattern occurs to further analyse it. The output must be a set of intervals where the instructor detected a change of activity and wants to delve into the possible reasons.

This step is generalisable to both synchronous / asynchronous courses since although the reasons of the patterns can vary in each type of course, the instructor only needs to observe the changes of activity in this point. However, a second possible approach for synchronous courses could be selecting intervals for each unit of the course (e.g., each week) and analyse them independently.

3.2.2 Step 2: Analysis of the social dimension

Once the time intervals for the analysis are defined, the instructor starts collecting the social variables for each interval. Some useful variables include: total number of messages, total number of responses (or ratio between number of responses and number of threads), total number of votes, total number of different learners who contribute in the forum, and response time between the first post in the threads and their first responses. The objective is that these social variables can be useful to understand possible MOOC issues. For example, very low response rates to messages could call teachers' attention to try to understand the possible reasons. Moreover, response times can be relevant since although the learner eventual-

ly receives a reply, if there is too much time between the question and the response, that learner might not know how to continue, get demotivated and drop the course.

After collecting the identified variables, the instructor decides (for the particular context of the MOOC) if the value of each variable is high (H), medium (M) or low (L). This task is perhaps one of the most subjective parts, but it is important to give teachers the power to define their own thresholds for each category because of several reasons. The most important one is that in each course the instructor may have a different idea of what the desirable use of the forum is. While there may be courses where instructors want learners to discuss and maybe foster a more collaborative learning, there may be others where the forum is a secondary tool and teachers do not expect as high participation as in the previous case. Moreover, as the contexts and number of learners are different, it is not possible to provide absolute reference values. This is also justified by the fact that even in the same course, participation may not be considered equally at different stages.

Next, the instructor could reflect on these variables to identify if there is any pattern associated to the social variables. For example, if the teacher finds that there is a high number of messages, but these posts come from a limited number of learners, there are no votes, there are few responses and the response time is high, it could be inferred that something is going wrong with the forum. In contrast, if there are lots of messages, with many responses and votes, from a high number of learners and with a short response time, it can be said that the forum is working well from the social point of view. Table 1 shows some combinations and their possible conclusions, although each case must be considered carefully by the instructor depending on the specific contexts. It is also noteworthy that this step can be done can be perform regardless the course is synchronous or asynchronous as the intervals are already selected in the previous step.

TABLE 1
EXAMPLE CASES FOR THE ANALYSIS OF THE SOCIAL DIMENSION
(STEP 2): HIGH (H), MEDIUM (M), LOW (L)

Variable	Case 1	Case 2	Case 3	Case 4
No. messages	H	H	L	M
No. responses	L	H	H	M
No. votes	M	M	H	L
No. different learners	L	H	M	L
Response time	M	M	L	H
SOCIAL (Conclusion)	Bad	Good	Probably good	Probably Bad

3.2.3   Step 3: Analysis of the sentiments dimension

The third step of the 3S methodology focuses on the sentiments dimension. This entails that messages for each interval selected in Step 1 are classified depending on their polarity. We can analyse different sentiments (e.g., frustration or happiness) and different scales can be used (e.g., continuous or categorical with just positive or negative). Let assume as an example that there is only one general emotion for students regarding a message with three possible values: positive, negative and neutral. In this case, the instructor can analyse the proportion of positive, negative and neutral messages and determine if sentiments are generally good, bad or if there is no specif-

ic pattern. The threshold can be determined by each teacher according to the context of the MOOC, although, in this case, it is easier to establish a relationship between the number of messages of each category and the findings of this step. Neutral messages may be excluded if desired to consider only the ratio between positive and negative messages. For example, if 5% of messages are positive, 40% negative and 50% neutral, it could be said that in general, sentiments are bad in the MOOC since there are many more negative messages than positive ones. This step, as the previous one, is independent on the format of the course (i.e., synchronous/asynchronous).

3.2.3   Step 4: Analysis of the skills dimension

The fourth step of the 3S methodology refers to the skills dimension. The aim is to identify which skills have more mentions in the interval of time to understand if there is any issue with them. To that end, we consider skills as keywords that indicates the concepts that learners should grasp to obtain a skill (e.g., the skill "sum" represents the ability of being able to sum numbers). Although there may not be direct relationship between difficulty and number of posts related to a certain skill, when there is too much debate about a certain skill, it could be a symptom that there are troubles to fully understand it. Hence, the instructor needs to visualise what skills appear more in the messages. In that visualisation, teachers need to decide how many skills predominate and this will be dependable on the context of the MOOC, although it is advisable to avoid large sets of skills to facilitate the analysis.

Similarly, if the instructor lacks any skill in the analysis and he thinks this skill should appear more, he can consider this skill separately and try to identify if there is any pattern that explains the low number of mentions. Particularly, in a synchronous course, it may be desirable to map skills to weeks of the course and check if the predominant skills are the ones that should be discussed, according to the contents of the week. In asynchronous courses, the instructor will focus more on the general trends of the skills and the possible issues with them.

3.2.3   Step 5: Conclusions for each interval

The fifth step aims to combine the information from the previous steps to reach some conclusions for each of the intervals. Inputs of this step are the conclusions obtained for the social and sentiments dimensions in steps 2 and 3, and the skills identified in step 4. Moreover, it is advisable to include data regarding near intervals from those initially selected to be able to analyse the later transitions (i.e., what happens after a selected interval). For example, if the interval is between May 1-5, it is recommended to provide information about the period between May 6-10 too.

The hardest part is identifying patterns for the input. This task is highly dependable on the context and should be assessed by each instructor. However, some general cases can be indicated as a guide (see Table 2). For example, in case 1, an instructor observes the activity is increasing and the conclusions of the social dimension are good. However, the sentiments are bad in the messages, which

are related to skills that should be learned in the interval considered (assuming a synchronous course). After analysing the later transition, it is observed that activity decreases as it was before the interval, and sentiments improve. When the instructor observes that, he may infer that new skill generated debate as learners did not understand them at first (and hence they were negative), but then they managed to understand them, which is good. This situation is generally acceptable as learners finally seem to have grasped the contents, which are challenging enough to generate debate. Nevertheless, the instructor could reach a different conclusion in a specific scenario and the format of the course may affect the results.

TABLE 2
EXAMPLE CASE FOR THE CONCLUSIONS FOR EACH INTERVAL (STEP 5): INTERVAL (INT), LATER TRANSITION (LAT), INCREASES (+), DECREASES (-), REMAINS (0), RELATED TO INTERVAL (REL)

Case	CASE 1		CASE 2		CASE 3	
	INT	*LAT*	*INT*	*LAT*	*INT*	*LAT*
Activity	+	-	-	+	+	-
Social	Good	Good	Bad	+	Bad	0
Sentiments	Bad	Get better	Very good	Good	Bad	0
Skills	REL	Any	REL	Any	Any	Any
Conclusion	Acceptable		Good		Bad	

In addition to these considerations, timing can also be taken into account for synchronous courses because a specific pattern could be related to a specific event of the course (e.g., deadline, release of materials, etc.). This can be extrapolated to any course when referring to actions taken by the instructors (e.g., reminder emails).

3.2.3 Step 6: Global conclusions

The last step consists on combining the results obtained in the previous step to get global conclusions of the course. For example, it may happen that in most of the intervals, the conclusion found is that learners' behaviour is affected by reminder emails. In this step, the instructor can review each interval to determine if there is any pattern that produces that or just conclude that reminder emails alter the forum activity. This will allow instructors to have a better perspective of the course and reflect on what was happening to improve future editions.

The drawback is that while previous steps can be done with partial data (teachers can download weekly data and analyse it), this step requires the whole information of the course, or at least a significant portion to reach the conclusions. However, findings obtained here may be more relevant as they are not tied to an interval of time.

In this methodology, step 1 must be taken once to select the intervals, while steps 2-5 need to be taken sequentially for each interval, and step 6 should be taken at the end. While step 6 permits to iterate again over the previous steps to identify if there is something missing that could be used to complete this last step. After completing the process, the instructor should be able to make a report on social activity based on the forum data.

4 PROPOSED VISUALISATION TOOL

In this section, the LATƎS (Learning Analytics Tool for the 3S methodology) visualisation tool for forum-related data is presented to support the proposed 3S methodolo-

gy. LATƎS has been designed and implemented for edX / Open edX since this is the most used open source MOOC platform. The design of LATƎS is described first and next each of the functionalities it supports.

4.1 Dataset and design considerations

Dashboards tools usually need some data to process. LATƎS is designed to cope with edX / Open edX data exclusively and the 3S methodology. Consequently, data used must have a common format to ensure the tool supports any kind of course deployed in edX / Open edX. Particularly, as the aim is to include forum-related interactions, the study focuses on the information related to the forums of the course, which is provided in a file (named with format *{org}-{course}-{run}-{site}.mongo*) that can be retrieved from the Database Data [35]

This file contains all forum messages and the characteristics of the course discussion interactions (e.g., votes, number of replies, timestamps, etc.). These interactions are different depending on the type of message, which can be of two types: *CommentThread* or *Comment*. A *CommentThread* is the first message of a thread, while a *Comment* is used for responses regardless their hierarchical level (i.e., response to a *CommentThread* or response to a response). It is noteworthy that currently a two-level hierarchy is supported, so *Comments* can be only first-level or second-level (i.e., response to a *CommentThread* or response to another *Comment)*.

Regarding the design of LATƎS, one important aspect has been the choice of the format. As MOOCs work on online Web platforms, users are used to access content and information related to MOOCs via web (e.g., edX Insights offers the visualisations in that format). Because of that, a Web application design has been chosen.

The development of LATƎS has been done using the Java Servlet Container Apache Tomcat[1], which provides a Web server to execute web applications. The main elements of the application are servlets and JSP (Java Server Pages) and the design is based on the MVC model (Model-View-Controller) [36], which allows separating the interactions with data (of the forum), the control of the application, and the graphical interface. LATƎS uses the following technologies: Java for the servlets which control the application, JSP (Java Server Pages) for the pages, CSS (Cascading Stylesheets) for the page style, JavaScript (including its visualisation libraries Google Charts[2] and D3.js[3]) for adding interactive elements and graphics in the pages, XML (eXtensible Markup Language) for the deployment descriptor of the application, and JSON (JavaScript Object Notation) for the data. Regarding the functionalities, the controller is in charge of collecting the data from the user (which is in the model component), managing users' actions in the application, and cleaning and processing the data with algorithms to provide the information to the view component, which handles the visualisations. A general architecture diagram with the components is shown in Figure 2.

---

[1] https://tomcat.apache.org/tomcat-9.0-doc/index.html
[2] https://developers.google.com/chart/interactive/docs/
[3] https://d3js.org/

4.2 Functionalities of LATƎS in connection with the 3S methodology

First, before the user selects any option, the main view is displayed (see Figure 3). This view is separated in three parts: the first one is used to upload the file which contains the forum interactions; the second one is used to define and visualise the skills that learners should develop within the course (defined by the instructor); the third one is used to select the different visualisations the tool provides (allowing filtering by time), which are all related to each one of the dimensions of the 3S methodology (the first four options are related with the social dimension, the fifth option with the sentiments dimension and the latter with skills dimension). Note that the aim of the tool is detecting general trends and visualisations are kept simple to achieve that goal while facilitating the understan...
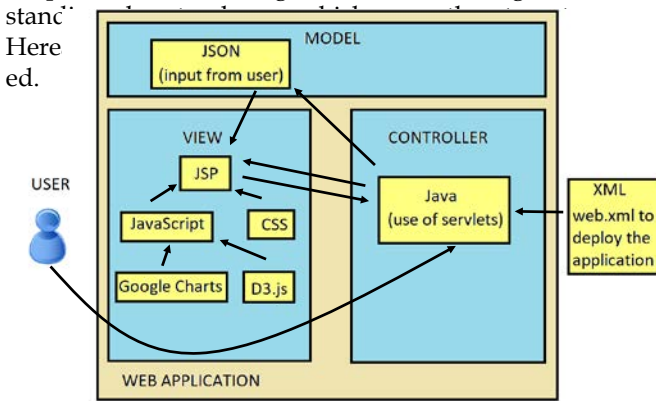
Here
ed.

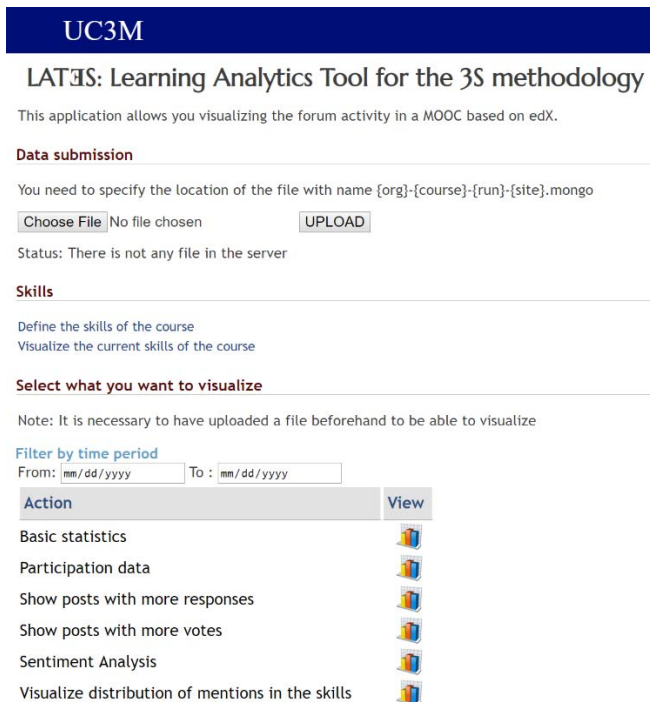Fig. 2. Architecture of the Web application

Fig. 3. Main view of the Web application

4.2.1. Data submission

In LATƎS, the first step to visualise data is uploading the

raw data containing the forum activity. As data is collected by edX, institutions cannot access the database directly, and they only receive data packages from edX periodically (a person designated as "data czar" by the institution receives the data packages). Because of that, teachers typically get a file with forum data from the data czar (once available), which can be uploaded/imported in the tool. To upload the data file, the option Choose File displays a window for selecting it. Once the file is selected, the path will appear on the screen and the uploading process starts when clicking UPLOAD button. The form used to upload data uses multipart/form-data encoding and admits a maximum size of 50 MB, which is reasonable considering that each post is estimated to have 1 KB, which entails that the tool could handle up to 50,000 messages. The upload process has been carried out with two external libraries: commons_fileupload-1.3[4] and commons-io-2.4[5]. When a file is uploaded, it is stored in a temporary directory to be used during the session. To avoid overloading files in the server, files are removed when the session expires, or when the user uploads another file, which substitutes the former one.

4.2.2. Time filtering

The second option is used to indicate a specific period of time to analyse data. When an instructor wants to visualise data, it is possible that he does not want to see the whole course, but a part of it; this is interesting to compare different parts of the course or to filter time periods for the analysis, as it is done in the 3S methodology (Section 3.2) once Step 1 is carried out. Because of that, two selectors have been defined to filter data by time. When each of them is clicked, a calendar is opened, and the user can indicate the dates. It is noteworthy that this filtering is optional. If no dates are given, all messages will be considered. If only one date is specified, the period will be only restricted by one side (i.e., from the beginning to date X or from date Y to the end).

4.2.3. Basic statistics (Social dimension)

Once data are on the server (whether time filtering is used or not), different visualisation options can be used. The first one provides some general information regarding the forum activity. This includes some indicators to be used in the social analysis, such as the total number of posts, threads and positive votes. Then, this functionality shows how many threads have at least one response. This value, together with the percentage of threads with response, gives an idea of how well the forum is working. Figure 4 shows a graphic with the forum activity per day. This graphic can be used for the step 1 of the 3S methodology.

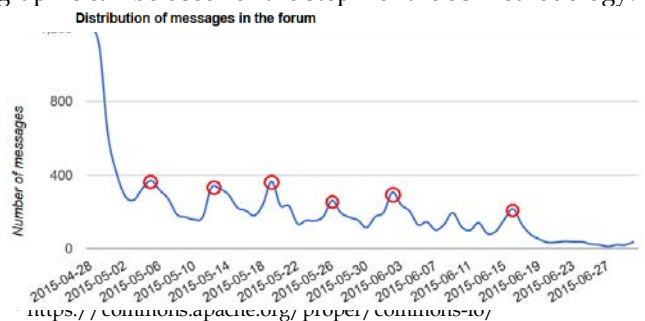https://commons.apache.org/proper/commons-io/

Date

Fig. 4. Distribution of messages in the forum over time in the example MOOC.

Apart from these statistics, there is an analysis of the time learners have to wait until their posts are answered. For this analysis, as the response time is the time between the learner posted the message which opens a thread and the first response, responses have been classified in three categories depending on its level (first or second) and position (according to the time when they were posted): (1) first responses (the oldest in time) to a thread, (2) first-level responses excluding first responses and (3) second-level responses (responses to first-level responses). From these categories, only the first one is relevant for the response time, although the others can give information about how much debate is in each thread.

The response time is shown with two metrics (the mean and the median). Moreover, a histogram has been implemented showing the response time of messages in hours (see Figure 5a). In this histogram, it is noteworthy that the last bar concentrates those messages with greater response times (to facilitate the representation), and that unanswered messages are always excluded (as there is no response time for them).

4.2.4. Participation data (Social dimension)

Participation in MOOCs is a well-known field of study because of the high variances between the number of enrolled users and the number of learners who actually participate, what produces very low completion rates (even lower than 10% [37]). In this case, as LAT3S is focused on analysing forum data, the aim is to show how many learners contribute in the forum and identify if they participate occasionally (e.g., only once typically to introduce themselves) or if they post messages regularly. Because of that, the number of messages to be in certain groups (best 1%, best 25% and best 50%) has been computed. Moreover, a histogram has been implemented (see Figure 5b), showing how many times learners have contributed in the forum (considering only those who contributed at least one to avoid a huge bar in "0").
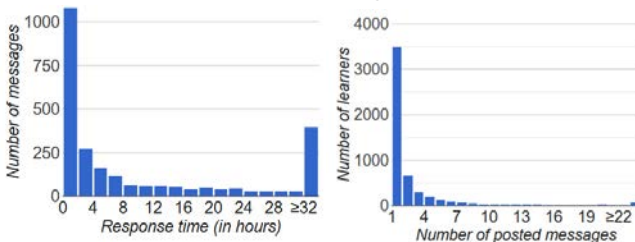

Fig. 5. (a) Histogram of response time (in hours) for threads in the forum in the example MOOC. (b) Histogram of participation of learners in the forum of the example MOOC

4.2.5. Posts with more responses/votes (Social dimension)

The next functionality aims to provide a list of the threads with the highest number of responses, so that teachers can get insights of what topics produce higher activity in the forum. This functionality is very similar to what edX offers, although, in the case of edX it is not possible to filter by time. This is an important limitation because if there are messages with a high activity in the first days of the course, they will remain on top in the latter weeks, while they may not be that relevant at those stages. Because of that, this functionality enhances the existing one, allowing time filtering.

When the user selects this functionality, a table with five columns is displayed. The first column indicates the order (e.g., "1" for the thread with more responses, and so on). The second column has several radio boxes, allowing users to select specific threads to visualise the whole conversation. The third column indicates the number of responses, the fourth one the title of the message, while the fifth one includes the first 50 characters of the message which opened the thread. Moreover, there is a button to redirect the page to another one with the detail (all messages of the thread) of the message selected.

Similarly, another functionality has been implemented for the number of votes. The idea is the same as the one already presented, but messages are sorted by the number of votes instead of by the number of responses. The view of the page is also equivalent, and only the meaning of the third column changes to indicate number of votes.

4.2.6. Sentiment Analysis (Sentiments dimension)

In the previous functionalities, all indicators and graphics could be obtained from the characteristics of the messages (e.g., votes, responses, etc.), but none of them analysed the text of the message. This functionality is different in that sense, and uses text mining techniques to determine the polarity of the messages (i.e., if a message is positive, negative or neutral), which can be used for the sentiments dimension of the 3S methodology.

To do that, a lexicon-based approach was used. This means that the polarity was obtained taking into account the polarity of words in the English lexicon, instead of using machine learning techniques. The method used is based on the one proposed by Hu and Liu [38], which mainly relies on the number of positive and negative opinion words (adjectives which express opinion). Particularly, there is an *orientation* variable which increases or decreases depending on whether there are positive or negative words, and it also considers the effect of negation words (e.g., not, no, etc.) and the meaning of emoticons. More details of the algorithm for sentiment analysis can be found in a related paper [39].

Regarding the results shown in LAT3S, there is an initial table which indicates the number of positive and negative posts detected in the period selected. Furthermore, the percentage of positive messages is shown, considering only positive and negative messages (excluding neutral ones). Then, there is a figure which includes a column chart to indicate the absolute number of positive, negative and neutral messages over time, and a line chart indicating the evolution of the percentage of positive messages (as defined previously) over time (see Figure 6). The line chart provides a better idea of the evolution of the sentiments, although the absolute numbers in the column

chart also indicate how representative each percentage is (e.g., the percentage is not representative with few posts).

4.2.7. Definition and distribution of mentions of course skills (Skills dimension)

The last functionality aims to identify the skills of the course that produce more debate in the forums, so they can be analysed in the skills dimension of the 3S methodology. In order to visualise those skills, the instructors need to provide them using the option "Define the skills of the course", which appears in Figure 3.

The definition of course skills has been implemented in the client-side (using JavaScript) and it is designed in an interactive way, so that users can add dynamically the skills and only after all of them are defined, they are sent to the server. As some skills can be subskills for other skills, it is possible to build a hierarchical tree with them, although plain trees are also acceptable. Independently of the option, the tree is shown in the tool and it is automatically updated when a new skill is added.

For adding new skills, there is a textbox to write the name of the new skill and a selector to choose the parent node. The tree is implemented with D3.js, so there is a JSON associated with the tree. If the user wants to modify the current tree, he can view the JSON, modify it, and import it directly in another textbox. Similarly, the user can just directly paste the JSON and load it, if he has it from a previous session.

Once the skills are defined, the user can save them. This process stores the tree in the session to be used in the application. Another option of LATƎS "Visualize the current skills of the course" allows viewing the saved tree and also downloading the JSON file. This can be useful because this JSON file could be loaded easily in a future session without having to fill in all the skills again.

After defining the skills, LATƎS can analyse the messages to visualise which ones appear more often in the forum. The result of that process is a circle graphic (see Figure 7, where the skills in the border of the circle are part of the abovementioned tree and they are used to obtain the graphic). This graphic includes a circle for each skill and there can be inner circles for the child nodes of the tree, depending on the structure of the defined tree (in this case, there are seven main skills with serveral subskills each). Size and colour represent the number of times a skill appears. Size represents the absolute value of each skill, taking into account that when a skill has others nested, the number of mentions will consider the whole subtree. Colour has a similar meaning, but initial values are assigned on leaf skills to prevent skills with more children from having more chances to be bigger. Among leaf skills, the one that has more appearances is coloured in red, while the one with less mentions is coloured in green, as it may show that there are no troubles with that skill (no one discussed about it). For the rest of the leaf skills, they are coloured with an intermediate RGB colour between red and green. For colouring the higher-level skills, the average of the RGB values of their children is used. Therefore, a higher-level skill needs many 'red' subskills to be 'red', and it is not enough to have one red

subskill, unlike size, where one subskill can contribute significantly to the total size of a higher-level skill.

Regarding the concept of 'appearance' or 'mention', there are also a couple of considerations. First, as a message can have the same skill repeated many times, to avoid the effect of the times the user repeats the word, only one appearance of skill per message is considered. Second, it may happen that a user uses a synonym for the skill, but not the skill itself. To cover this situation, messages are scanned to look for skills or for synonyms of the skills (considering that once found the word or the synonym, there cannot be repetitions). Thus, if a synonym is found, it will be considered like the original skill. These synonyms are identified using the database WordNet [40], which includes groups of words with similar meaning. Moreover, the library JAWS[6] (Java API for WordNet Searching) has been used to look for synonyms in WordNet. Nevertheless, the instructor should note that synonyms sometimes can be common English words (e.g., verbs) that can alter the visualisation. Because of that, instructors should reflect carefully the skills they choose to avoid the possible linguistic limitations. Finally, in terms of implementation, it is noteworthy that the circle graphic is implemented with D3.js, which allows animations in the graphic (it is possible to enter inner circles) and has JSON file with the information to be displayed.
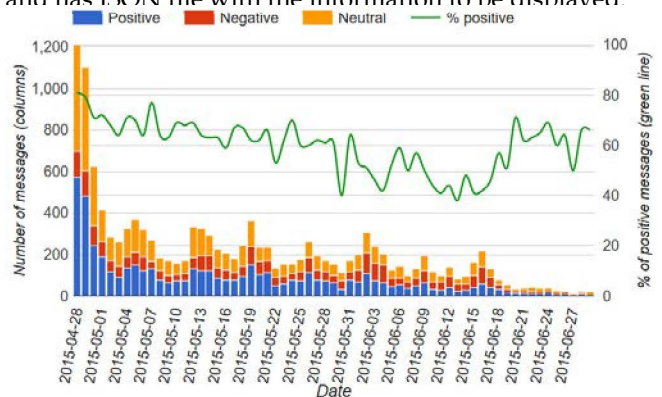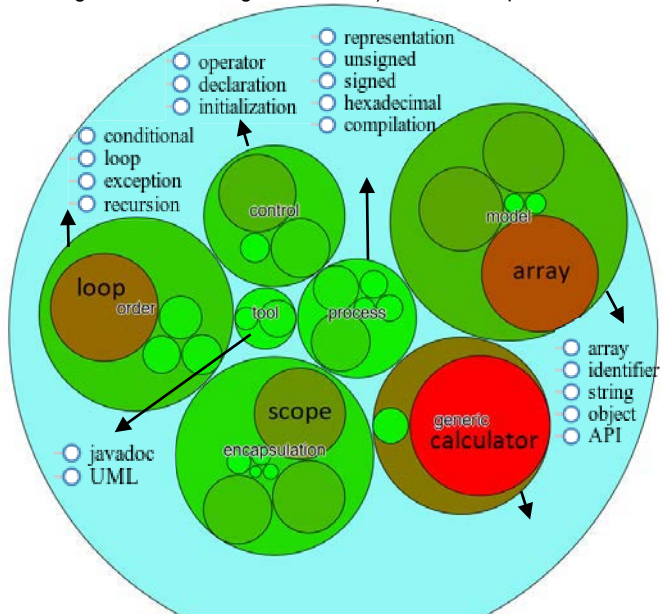

Fig. 6. Distribution of sentiments and percentage of positive messages in the forum posts over time (including positive and negative messages, and excluding neutral ones) for the example MOOC
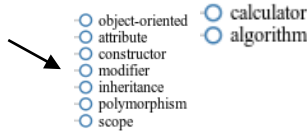
Fig. 7. Circle graphic representing the mentions of the skills of the example MOOC. Skills of each category are indicated in the border.

5 RESULTS AND DISCUSSION

This section illustrates the application of the 3S learning analytics methodology in a real MOOC about Java Programming, using LATƎS. First, a description of the MOOC is provided. Next, the details of each step of the 3S methodology are described for the case study.

5.1 Case study

This study was conducted using the data of the first edition of a MOOC about the basics of Java programming. The course, taught in English, was organised in five weeks, where learners had to watch video lectures and do activities to understand the contents, and then, they had to take weekly assignments. The methodology of this course was instructor-led. This means that there were dates for the release of materials and other dates for completing each of the assignments. The first date of the course was April 28, 2015, where the course begun, and the first deadline (for week 1) was three weeks later. Two weeks later (May 12), materials of the second week were released and from that date on, new materials were released in each successive week, and there were two weeks for completing each assignment.

In the course, there were two types of assignments: closed-ended and open-ended. Each week had a graded test, which consisted on closed-ended questions (e.g., multiple-choice questions, checkbox questions, numerical input questions, etc.). These tests weighted 15% of the final grade each. In addition, there were open-ended programming assignments on weeks 3 and 5, which counted 10% and 15%, respectively. The assessment method for programming assignments was peer-review, so learners had to implement their code and assess the code of their coursemates, based on a scoring rubric. The passing grade of the course was 60% although it was not necessary to pass (or even participate) in all the assignments to receive a passing grade. In the course, there were 95,555 enrolled users, but only 24,055 played at least one video (25.2%), only 16,317 submitted at least one ungraded exercise (17.1%) and only 1,507 passed the course (1.6%).

5.2 Application of the 3S methodology

As the steps of the methodology can be taken sequentially, this section details the findings for each of them for this case study.

5.2.1 Step 1: Selection of relevant intervals of time

The first step consists on selecting the intervals of time where there is a relevant trend change in forum activity. To do this step, it is possible to use the graphic provided by the tool regarding the number of messages over time (see Figure 4). If the first days of the course (when there is an exponential decrease of activity between April 28 to May 2) are excluded, then several peaks can be observed (marked in red). These intervals, which are the output of the first step are: (**1) May 03-08, (2) May 11-15, (3) May 17-21, (4) May 25-28, (5) June 01-04, and (6) June 14-18**.

5.2.2 Step 2: Analysis of the social dimension

The second step needs to consider each of the six intervals selected in Step 1 and analyse the value of the social indicators, which can be gathered from the tool. The number of messages and response times are directly provided in the option "Basic statistics". Then, as for the responses, it is possible to report the absolute number, but LATƎS also offers the percentage of messages with response. So, in this case, both values are considered for the analysis. As for votes, both absolute number of votes and the average of votes per message are used. Finally, the number of different learners who contribute to the forum has been taken from the "Participation data" option.

For each of the indicators, values have been obtained (see Table 3). Next, these values have been compared to other values in the course, and classified (in a subjective way) in high (H), medium (M) or low (L). For the case of number of messages, as all intervals are positive peaks, they have been considered 'high'. Afterwards, a "social" conclusion has been drawn for each interval in terms of GOOD/FAIR/BAD. For example, the first interval has been considered 'bad' because there are many messages, but the number of responses and votes is high, and the response time is higher than in other intervals. Moreover, the number of learners is high, what means that many learners posted in the forum, but their messages were mostly unanswered, which is an undesired behaviour. In contrast, interval 4 shows that there are many messages, responses and votes, but few learners posting. This may entail that learners wrote their messages and these posts received replies, which were relevant in several cases as the number of votes was also high. Because of that, this interval has been classified as 'good'.

TABLE 3
VALUES OF THE INDICATORS USED IN THE SOCIAL DIMENSION
FOR EACH INTERVAL (I)

Variable	I1	I2	I3	I4	I5	I6
No. messages	1733 H	1341 H	1252 H	800 H	944 H	680 M
No. responses Messages with response	605 231 /1128 21% L	581 253 /760 33% M	591 219/ 661 33% M	369 143/ 431 33% M	620 182/ 324 56% H	424 127/ 256 50% H
No. of votes Average votes/message	227 0.15 L	289 0.25 M	300 0.28 M	205 0.30 H	245 0.37 H	84/485 0.17 L
No. of different learners	1109 H	733 H	596 M	386 M	359 L	249 L
Response time	9.03h M	7.60h M	8.93h M	6.28h L	5.99h L	7.18h M
SOCIAL (Conclusion)	BAD	FAIR	FAIR	GOOD	GOOD	FAIR

5.2.3 Step 3: Analysis of the sentiments dimension

The third step can be taken using the "Sentiment Analysis" option of LATƎS, which reports the number of positive, negative and neutral messages over time. Particular-

ly, these values are computed for each interval (from those selected in Step 1), together with the percentage of positivity (relationship between positive messages over the total of positive and negative posts, excluding neutral ones). Once these values are available, the output of this step is a classification for each interval in GOOD/FAIR/BAD. The category has been assigned based on the percentage of positivity taking into account the values of all intervals and the overall value of 64.33%. Results are provided in Table 4.

TABLE 4
VALUES OF THE INDICATORS USED IN THE SENTIMENTS DIMEN-
SION FOR EACH INTERVAL (I) (STEP 3)

Variable	I1	I2	I3	I4	I5	I6
Positive posts	702	537	540	331	313	199
Neg. posts	314	280	292	208	328	238
Neutral posts	717	524	420	261	303	243
% of positivity	69%	66%	65%	61%	49%	46%
SENTIMENTS (Conclusion)	GOOD	FAIR	FAIR	FAIR	BAD	BAD

5.2.4 Step 4: Analysis of the skills dimension

In the fourth step, the last functionality of LATƎS ("Visu-alise distribution of mentions in the skills") was used to discover what skills generated a higher debate in each interval. Table 5 shows the results obtained with the col-ours indicating the degree of mentions (in quantity) for each skill (from those which appear in Figure 7, which have been elaborated based on ontologies about Java language [41][42]). It is interesting to analyse that in the first interval, there were a high number of messages relat-ed to "calculator" and "operator", which were covered in Week 1, the week where the interval was framed. The second interval corresponds also to Week 1, but the mate-rials of Week 2 were released the second day of the inter-val. In this case, results show that previous skills remain but "loop" and "array" (covered in Week 2) stand out, which means that learners studied the materials in the first days (otherwise, they would not probably know what arrays or loops were).

TABLE 5
SKILLS WITH MORE MENTIONS FOR EACH INTERVAL (I) (STEP 4)

I	Skills
I1	Calculator (red) and operator (olive)
I2	Loop (red), array (brown), calculator (brown) and operator (olive)
I3	Array (red), loop (red), calculator (brown), string (brown) and scope (brown)
I4	Array (red) and loop (olive)
I5	Array (red), loop (red), object (brown), constructor (olive) and Javadoc (olive)
I6	Array (red), constructor (red), object (red), attribute (red), scope (red), declaration (red) and compilation (brown)

A similar pattern occurs in the third interval. This in-terval includes the last days before the first deadline and the release of the materials of Week 3. In this case, skills related to Week 2 still predominate, unlike in the previ-ous case, but new skills also arise (e.g., "scope"). Fur-thermore, skills related to Week 2 seem to appear throughout the course since they are those with stand out in intervals 4 and 5 as well. Nevertheless, interval 5 also

shows the appearance of skills related to object-oriented programming (e.g., "object" and "constructor"), which are covered in Week 4. However, the contents of Week 4 are released during Interval 4. This entails that while learners are more likely to go through the new contents since they are released at the beginning, they usually engage with them later in the final stages of the course. Finally, interval 6 shows a wide variety of skills. While half of them are related to object-oriented programming, other words such as "compilation" arises. A possible reason for the last one is that there is a programming assignment in the interval and learners could face diffi-culties to make it work.

5.2.5 Step 5: Conclusions for each interval

The fifth step consists of putting results from previous steps together to get some conclusions. Additionally, intervals are compared with the transitions to see how the evolution is. The transition of each interval has been de-fined as the next period of the interval, with the same days as the interval to make fair comparisons. Table 6 summarises the results of this analysis. In that table, for each interval (I), the results from previous steps appear (in skills only the number of predominant skills appear), whereas for each transition (T), there is a '+' when there is an improvement with respect to the interval, a '-' when things get worse, or a '0' when the situation is equivalent. For the case of skills, those symbols indicate if there are more or less predominant skills.

From the six intervals, the most representative ones are intervals 2 and 5. In interval 2, the social and sentiments dimensions are fair in the interval, but both get worse afterwards, while the predominant skills remain. This may entail that learners had more difficulties (with the same contents) as their sentiments decreased and the forum was not as active as previously to be a way to solve the problem, which is something bad. To corroborate the actual difficulties, the messages with more debate were reviewed. It was found that they were related to problems with calculations (the first one starts with "I really don't understand this new topic"), arrays syntax or using loops to traverse arrays which all belong to the identified skills. The recommendation in this case may be adapting the materials that explain the skills to see if there is an im-provement in the following run of the MOOC.

TABLE 6
SUMMARY OF THE CONCLUSIONS FOR EACH INTERVAL (I) (STEP 5): GOOD (G), FAIR (F), BAD (B)

Interval	I1		I2		I3		I4		I5		I6	
	I	T	I	T	I	T	I	T	I	T	I	T
Activity	+	-	+	-	+	-	+	-	+	-	+	-
SOCIAL	B	+	F	-	F	0	G	-	G	0	F	-
SENTIMENTS	G	-	F	-	F	-	F	-	B	+	B	+
SKILLS	2	+	4	0	5	-	2	0	5	+	7	-
CONCLUSION	Fair		Bad		Fair		Fair		Good		Fair	

Interval 5, in contrast to interval 2, shows that there was a good social activity, although sentiments were bad. However, after the interval, the social dimension was also good, but sentiments improved. This might mean that the contents were challenging enough to generate debate, but

the forum worked properly as a medium of support, and finally learners managed to solve their problems. In this case, perhaps the difficulty was high as sentiments were too negative compared to other weeks, but after the interval, sentiments improved considerably, which is good. This can be corroborated with messages, as one of the messages with generated more debate was called "Opinions about the difficulty of the exam 3", which criticised the difficulty of the test of that week. Apart from that, the evaluation of peer review activity in this week was found to cause frustration among learners.

Another important observation, which can explain most of the intervals, is that peaks of activity corresponds to deadlines of dates when activities are released, which implies that these dates can activate the forum. Moreover, instructors usually posted reminder messages in those dates, which could also affect the increase of activity. The only exception was in the first peak (May 05), which does not correspond to any deadline, but that date is just one week after the start of the course. This fact may suggest that there were some learners who started the course after one week and started posting in the forum.

5.2.6 Step 6: Global conclusions

After reflecting on data from previous steps, the findings of the case study (from the last step) are the following:

1. Learners tend to be more active after instructors' actions. The distribution of messages over time showed significant positive peaks of activity (the activity increased between 50-130% in the periods of those peaks) when there was a deadline, a release of materials or reminder e-mails by instructors. This means that learners usually react to instructors' actions (it is an action-reaction process).

2. Not many learners use the forum. Although social information indicated moments where learners replied more, voted more, etc., participation data showed that 5,126 learners contributed at least once in the forum, which is 21.31% of the learners who watched at least a video. Considering that 67.81% of learners participated only once and there were 3,782 posts in the introductions thread, it seems that the participation was not so high. However, this is not necessarily a problem since there are learners who can be engaged but not participate in the forum. In fact, 49.97% of learners who passed did not post any messages.

3. Social activity was generally acceptable among learners who participate in the forum. In most of the intervals, the analysis of the social dimension showed that there were over 200 messages per day and learners received responses. In this case, it is not possible to reach 100% of threads with responses because some of them are just comments (e.g., when a learner introduces himself), but values are acceptable, though some intervals could be improved. Apart from that, response times were also good since, in average, messages were replied in less than a day, which sounds reasonable.

4. Learners positivity decreases over time and particularly before deadlines of open-ended assignments. The evolution of sentiments over intervals shows that users were more positive at the beginning of the course and their positivity was dropping progressively, with prominent peaks three days before of the programming assignment. The fact that deadlines were on Tuesdays also gave the conclusion that learners tended to use the weekends for completing MOOC activities as they could be more frustrated when they faced the assignments on Saturdays and did not get quick answers.

5. There are some skills that entail difficulties in the whole course. While the analysis of skills reveals that new skills become relevant as the course evolve and new contents are released, it is also identified that some skills like "arrays" or "loop" are present during the whole course. In addition, interval 2, which corresponds to the period where these skills should be learnt, showed undesired (bad) behaviour of the forum. Hence, it should be recommendable to give supporting material for those skills. In contrast, other possible skills like polymorphism generated very few mentions. Therefore, instructors could incentive the debate around those skills.

6. Difficulty of the course should be reviewed at certain points. There are some points (e.g., week 3) where the difficulty could be high for learners, so it should be analysed if there should be any adaptation to ensure learners can complete the materials properly if they watch the lectures and do the exercises.

6 CONCLUSIONS

This work has presented the 3S methodology to analyse forum activity in MOOCs, following six steps. To illustrate its use, a visualisation tool, LATƎS, was developed, which included several functionalities regarding each of the dimensions of the methodology (social, sentiments and skills). LATƎS was then used with a MOOC about Java programming to show how the steps of the methodology could be applied in a real context.

The application of the 3S methodology showed that it can support instructors to understand learners' behaviours such as the increase of activity near critical dates. Teachers can also benefit by increasing their awareness on troubles to acquire certain skills, which may lead to the re-design of some parts or the course. Nevertheless, the 3S methodology should be adapted to each course, regardless the platform or if the MOOC runs in synchronous or asynchronous modes.

In this paper, we focused on the platform edX / Open edX to guide the 3S methodology and also complement the visualisations already offered, which mainly lack forum information. We gathered raw data from that platform and obtained several indicators through the transformation of the initial data. This allowed implementing functionalities regarding the participation of the course (focused on the time and number of learners), the response times, the messages with more votes or responses, the sentiments users showed with their posts and the skills that produced a higher debate in the course.

Despite being able to offer a methodology to analyse the data, this work is not exempt of some limitations that are worth mentioning. The most important one is that the methodology has only been applied with one course as an

example, and more cases should be required to evaluate and refine it. Moreover, the evaluation of LAT3S has been limited and there should be a further study of usability aspects with the evaluation of different users to ensure the tool is useful and easy to use. In addition, we tried to give some guidelines about how to interpret the data but the intervention of teachers evaluating the specific context is necessary. Furthermore, the tool can be only applied with English courses due to the dictionary used for sentiment analysis and it is not real-time because of the edX policies with sharing data.

Because of these limitations, as future work, it would be interesting to use the 3S methodology with other courses and provide more specific tips about how to interpret data in different situations (and automatic diagnosis if possible). Moreover, the analysis of those courses will allow understanding new learning scenarios and comparing the conclusions obtained with the ones obtained here. This would be relevant as the instructional conditions can vary the results considerably [43]. Among the target courses, it would be interesting to analyse other thematic areas, because the behaviour of those learners may be very different. Furthermore, it should be relevant to replicate the methodology with asynchronous courses to understand their differences with synchronous courses.

Apart from that, a possible future work would be enabling the automatic skills extraction from course materials (e.g., text of video scripts) and map them with the forum messages to make the analysis easier and quicker, although teachers should keep the possibility to add skills manually if desired. Moreover, the functionality could be extended to differentiate synonyms of a word in other contexts. Additionally, the sentiment analysis tool could be refined by analysing new algorithms which could enhance the predictive power, by considering the addition of positive/negative words by the instructor and including dictionaries in other languages. In addition, new functionalities could be added to alert individual students whose behaviour is not as desired, and Social Network Analysis (SNA) could be considered. Finally, as for the evaluation of LAT3S, the tool could be used while a course is being developed to check if it can support detecting troubles and provide useful information regarding the learning processes to help learners acquire knowledge better.

ACKNOWLEDGMENT

REFERENCES

[1] B. Grainger, "Massive Open Online Course (MOOC) Report 2013," *University of London*, 2013.

[2] C. Alario-Hoyos, M. Pérez-Sanagustín, C. Delgado Kloos, M. Muñoz-Organero, and A. Rodríguez-de-las-Heras, "Analysing the impact of built-in and external social tools in a MOOC on educational technologies," *Proc. 8th Eur. Conf. on Technol. Enhanced Learn.*, pp. 5-18, Sep. 2013.

[3] S.F.J. Mak, R. Williams, and J. Mackness, "Blogs and Forums as Communication and Learning Tools In a MOOC," *Proc. Int. Conf. on Networked Learn.*, pp. 275-285, May 2010.

[4] J. Bueder, C. Schwind, A. Rudat, and D. Bodemer, "Selective reading of large online forum discussions. The impact of rating visualizations on navigation and learning," *Comput. Hum. Behavior*, vol. 44, pp. 191-201, Mar. 2015.

[5] S. Few, *Information Dashboard Design: The Effective Visual Communication of Data*. Sebastopol, Calif.: O'Reilly Media, 2006.

[6] J. Grann and D. Bushway, "Competency map: visualizing student learning to promote student success," *Proc. Int. Conf. Learn. Anal. Knowl.*, pp. 168-172, Mar. 2014.

[7] edX, "Using edX Insights", *edX*, http://edx.readthedocs.io/projects/edx-insights/en/latest/. 2018.

[8] A. Peña-Ayala, *Learning Analytics: Fundamental, Applications, and Trends: A view of the Current State of the Art to Enhance e-Learning*. Cham, Switzerland.: Springer, 2017.

[9] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J.L. Santos, "Learning Analytics Dashboard Applications," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1500-1509, Feb. 2013.

[10] B.A. Schwendimann, M.J. Rodríguez-Triana, A. Vozniuk, L.P. Prieto, M.S. Boroujeni, A. Holzer, D. Gillet, and P. Dillenbourg, "Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 30-41, Jan. 2017.

[11] M. Pérez-Sanagustín, C. Alario-Hoyos, and C. Delgado Kloos, "FLINN: A Framework to Characterize Technology Enhanced Formal, Non-Formal and Informal Learning Situations," *Int. J. Human Capital Inform. Technol. Professionals*, vol. 5, no. 2, pp. 38-51, Apr.Jun. 2014.

[12] K. Verbert, S. Govaerts, E. Duval, J.L. Santos, F.V. Assche, G. Parra, and J. Klerkx, "Learning Dashboards: an Overview and Future Reserach Opportunities," *Personal and Ubiquitous Computing*, vol. 18, no. 6, pp. 1499-1514, Aug. 2014.

[13] Y-C. Yu, D.Y. Shing-chern, and D-R. Tsai, "Social interaction feedback system for smart classroom," *Proc. IEEE Int. Conf. Consumer Electronics*, pp. 500-501, Aug. 2012.

[14] L.H.S Do. "Supporting Reflection and Classroom Orchestration with Tangible Tabletops," PhD dissertation, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, 2012.

[15] L. Ali, M. Hatala, D. Gašević, and J. Jovanović, "A qualitative evaluation of evolution of a learning analytics tool," *Comput. Edu.*, vol. 58, no. 1, pp. 470-489, Jan. 2012.

[16] J.A. Ruipérez-Valiente, P.J. Muñoz-Merino, D. Leony, and C. Delgado Kloos, "ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform," *Comput. Hum. Behavior*, vol 47, pp. 139-148, Jun.

2015.

[17] C. Shi, S. Fu, Q. Chen, and H. Qu, "VisMOOC: Visualizing video clickstream data from Massive Open Online Courses," *Proc. IEEE Pacific Visualization Symp.*, pp. 159-166, Apr. 2015.

[18] J.A. Ruipérez-Valiente, P.J. Muñoz-Merino, J.A. Gascón-Pinedo, and C. Delgado Kloos, "Scaling to Massiveness With ANA-LYSE: A Learning Analytics Tool for Open edX," *IEEE Trans. Human-Machine Syst.*, vol. 47, no. 6, pp. 909-914, Jan. 2017.

[19] S.B. Shum and R. Ferguson, "Social Learning Analytics," *J. of Edu. Technol. & Society*, vol. 15, no. 3, pp. 3-26, Jul. 2012.

[20] C. Alario-Hoyos, M. Pérez-Sanagustín, C. Delgado Kloos, H.A. Parada G., and M. Muñoz-Organero, "Delving into Participants' Profiles and Use of Social Tools in MOOCs," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 260-266, Jul.-Sep. 2014.

[21] G. Chen, D. Davis, C. Hauff, and G-J. Houben, "On the Impact of Personality in Massive Open Online Learning," *Proc. Conf. on User Modeling Adaptation Personaliz.*, pp. 121-130, Jul. 2016.

[22] D. Yang, R.E. Kraut, and C. Rosé, "Exploring the Effect of Student Confusion in Massive Open Online Courses," *J. Edu. Data Mining*, vol. 8, no. 1, pp. 52-83, 2016.

[23] S. Chaturvedi, D. Goldwasser, and H. Daumé III, "Predicting Instructor's Intervention in MOOC forums," *Proc. Ann. Meeting Assoc. Comput. Linguistics*, pp. 1501-1511, 2014.

[24] A. Ramesh, S.H. Kumar, J.R. Foulds, and L. Getoor, "Weakly Supervised Models of Aspect-Sentiment for Online Course Discussion Forums," *Proc. Ann. Meeting Assoc. Comp. Linguistics*, pp. 74-83, Jul. 2015.

[25] A. Bakharia, "Towards Cross-domain MOOC Forum Post Classification", *Proc. ACM Conf. Learning@ Scale*, pp. 253-256, Apr. 2016.

[26] P.M. Moreno-Marcos, P.J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, and C. Delgado Kloos, "Analysing the predictive power for anticipating assignment grades in a Massive Open Online Course," *Behaviour & Information Technology* (In Press)

[27] E. Hoque and G. Carenini, "ConVis: A Visual Text Analytic System for Exploring Blog Conversations;" *Comput. Graphics Forum*, vol. 33, no. 3, pp. 221-230, Jul. 2014.

[28] H. Qu and Q. Chen, "Visual Analytis for MOOC Data," *IEEE Comput. Graph. Applic.*, vol. 35, no. 6, pp. 69-75, Nov-Dec. 2015.

[29] S.M. Jayaprakash, "SNAPP (Social Network Analysis & Pedagogical Practices) for Sakai CLE v2.8x & v2.9x)," 2014. [Online]. Available: https://confluence.sakaiproject.org/pages/viewpage. action?pageId=84902193.

[30] M. Chitsaz, L. Vigentini, and A. Clayphan, "Towards the development of a dynamic dashboard for FutureLearn MOOCs: insights and directions," *Proc. Int. Conf. Innovation, Practice and Research Use of Educ. Technol. Tertiary Educ.*, pp. 116-121, Nov. 2016.

[31] M. Klüsener, and A. Fortenbacher, "Predicting students' success based on forum activities in MOOCs," *Proc. Int. Conf. Intell. Data Acquisition Adv. Comput. Syst.: Technol. Appl.*, vol. 2, pp. 925-928, Sep. 2015

[32] B. Grawemeyer, M. Mavrikis, W. Holmes, A. Hansen, K. Loibl, and S. Gutiérrez-Santos, "The impact of feedback on students' affective states," *Proc. Work. Int. Con. Artif. Intellig. Edu.*, vol. 7, pp. 4-13, Jun. 2015.

[33] X. Wang, D. Yang, M. Wen, K. Koedinger, and CP. Rosé, "Investigating How Student's Cognitive Behavior in MOOC Discus-sion Forums Affect Learning Gains," *Proc. Int. Conf. on Edu. Data Mining*, pp. 226-233, Jun. 2015.

[34] D.F.O. Onah, J.E. Sinclair and R. Boyatt, "Exploring the Use of MOOC Discussion Forums," *Proc. London Int. Conf. Edu.*, pp. 1-4, Nov. 2014.

[35] edX, "EdX Research Guide Release," 2018. [Online]. Available: https://media.readthedocs.org/pdf/devdata/latest/devdata.pdf.

[36] G.E. Krasner and S.T. Pope, "A Description of the Model-View-Controller User Interface Paradigm in the Smalltalk-80 system," *J. Object Oriented Programming*, vol. 1, no. 3, pp. 26-49, 1988.

[37] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, "Attrition in MOOC: Lessons learned from drop-out students", *Int. Work. Learn. Technol. Edu. Cloud*, pp. 37-48, Sep. 2014.

[38] M. Hu and B. Lui, "Mining and Summarizing Customer Reviews," *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data*, pp. 168-177, Aug. 2004.

[39] P.M. Moreno-Marcos, C. Alario-Hoyos, P.J. Muñoz-Merino, I. Estévez-Ayres, and C. Delgado Kloos, "Semtiment Analysis in MOOCs: A case study," *Proc. IEEE Global Engin. Edu.*, pp. 1489-1496, Apr. 2018

[40] G.A. Miller, "WordNet: a Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, Nov. 1995.

[41] A. Kouneli, G. Solomou, C. Pierrakeas, and A. Kameas, "Modeling the Knowledge Domain of the Java Programming Language as an Ontology," *Proc. Conf. on Web-Based Learn.*, pp. 152-159, Sep. 2012.

[42] M-C. Lee, D.Y. Ye, and T.I. Wang, "Java Learning Object Ontology," *Proc. Int. Conf. Adv. Learn. Technol.*, pp. 538-542, Jul. 2005.

[43] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success", *Internet Higher Edu.*, vol. 28, pp. 68-84, Jan. 2016.

Pedro Manuel Moreno-Marcos is a PhD student at the Department of Telematics Engineering at the Universidad Carlos III de Madrid (UC3M). He received his B.S in Telecommunications Technologies Engineering in 2015 and his M.S Degrees in Telecommunication Engineering and Telematic in 2017 at UC3M. His areas or research interest include learning analytics, Educational Data Mining and MOOCs.

Carlos Alario-Hoyos is a Visiting Associate professor in the Department of Telematics Engineering at the Universidad Carlos III de Madrid. He received M.S. in Telecommunication Engineering in 2007 and PhD in Information and Communication Technologies from Universidad of Valladolid in 2012. Carlos has received several awards for his research, he is author of more than 60 scientific publications and has participated in more than 20 research projects. His skills and experience include research and development in MOOCs, social networks, collaborative learning, and evaluation of learning experiences, among others.

Pedro J. Muñoz-Merino is Associate Professor at the Universidad Carlos III de Madrid (UC3M), where he got his PhD in Telematics Engineering in 2009. Pedro has received several awards for his research. He is author of more than 90 scientific publications and has participated in more than 30 research projects. His skills and experience include research and development in learning analytics, educational data mining, evaluation of learning experiences, and gamification, among others.

Iria Estévez-Ayres is Associate Professor at the Telematic Engineering Department at Universidad Carlos III de Madrid (UC3M). She obtained her Telecommunication Engineering degree from Universidad de Vigo in 2001 and her PhD from UC3M in 2007. She has been involved in several research projects and is author in top international journals and conferences in the fields of technology-enhanced learning, networking, distributed systems and real-time systems.

Carlos Delgado Kloos received the PhD degree in Computer Science from the Technische Universitat Munchen and in Telecommunications Engineering from the Universidad Politecnica de Madrid. He is full professor of Telematics Engineering at the Universidad Carlos III de Madrid, where he is the director of the GAST research group, director of the UNESCO Chair on "Scalable Digital Education for All", and vicepresident for Strategy and Digital Education.