Link sito dell'editore: https://www.journals.elsevier.com/computers-in-human-behavior

Link codice DOI: https://doi.org/10.1016/j.chb.2018.04.033

Citazione bibliografica dell'articolo:

Elia, G., Solazzo, G., Lorenzo, G., Passiante, G. (2019) Assessing learners' satisfaction in collaborative online courses through a big data approach. *Computers in Human Behavior*, 133, 279-286

Assessing Learners' Satisfaction in Collaborative Online Courses through a Big Data approach

Gianluca ELIA (*)
Dept. of Engineering for Innovation, Campus Ecotekne, University of Salento, Lecce (Italy), gianluca.elia@unisalento.it

Gianluca SOLAZZO
Dept. of Engineering for Innovation, Campus Ecotekne, University of Salento, Lecce (Italy), gianluca.solazzo@unisalento.it

Gianluca LORENZO
Dept. of Engineering for Innovation, Campus Ecotekne, University of Salento, Lecce (Italy), gianluca.lorenzo@unisalento.it

Giuseppina PASSIANTE
Dept. of Engineering for Innovation, Campus Ecotekne, University of Salento, Lecce (Italy), giuseppina.passiante@unisalento.it

(*) Corresponding author

Abstract

Monitoring learners' satisfaction (LS) is a vital action for collecting precious information and design valuable online collaborative learning (CL) experiences. Today's CL platforms allow students for performing many online activities, thus generating a huge mass of data that can be processed to provide insights about the level of satisfaction on contents, services, community interactions, and effort. Big Data is a suitable paradigm for real-time processing of large data sets concerning the LS, in the final aim to provide valuable information that may improve the CL experience. Besides, the adoption of Big Data offers the opportunity to implement a non-intrusive and in-process evaluation strategy of online courses that complements the traditional and time-consuming ways to collect feedback (e.g. questionnaires or surveys). Although the application of Big Data in the CL domain is a recent explored research area with limited applications, it may have an important role in the future of online

education. By adopting the design science research methodology, this article describes a novel method and approach to analyse individual students' contributions in online learning activities and assess the level of their satisfaction towards the course. A software artefact is also presented, which leverages Learning Analytics in a Big Data context, with the goal to provide in real-time valuable insights that people and systems can use to intervene properly in the program. The contribution of this paper can be of value for both researchers and practitioners: the former can be interested in the approach and method used for LS assessment; the latter can find of interest the system implemented and how it has been tested in a real online course.

Keywords: Big Data; Clustering; Collaborative Learning; Learning Analytics; Learning Satisfaction; Sentiment Analysis.

1. Introduction

Today learning processes are networked and collaborative rather than isolated and bounded (Elia and Poce, 2010). They are held in technology-enhanced environments, which contribute to stimulate cognitive abilities and support knowledge construction (Kirschner et al., 2015). Therefore, collaborative learning (CL) arises as a strategic process to support lifelong learning initiatives within corporations, as well as to renew the educational offering of academic institutions. It relies on constructivism and connectivism (Vygotsky, 1978; Siemens, 2004; Conole et al. 2010; Ravenscroft, 2011) to promote active learning (Barkley et al., 2004) and create social spaces where participants can learn together, create new knowledge, share experiences, and develop competencies (Kreijns et al., 2003; Elia et al., 2009; Su et al., 2010). Monitoring and assessing learners' satisfaction (LS) in CL environments is crucial since it represents a key component that may influence the level of motivation and engagement of learners (Donohue and Wong, 1997), and reduce early drop-outs (Bolliger, 2004; Rigou et al., 2004).

Online learners perform many online activities, leaving significant streams of data on the web and generating very large data sets (Koedinger et al., 2010). Thanks to the increase of computing power and the reduction of costs for data processing and data storage, organizations can exploit the potential value embedded within these data sources (Pearson and Wegener, 2013). In particular, the emerging paradigm of Big Data that includes software development frameworks (e.g. Hadoop / Map Reduce / HBase), technological infrastructure and data management frameworks (e.g. Hadoop, Apache Flume, Elastic Search), and platforms for data analytics and visualization (e.g. RapidMiner, Kibana) allows for turning high volumes of fast-moving and diverse data into meaningful insights.

In the context of the present work, the term Big Data refers to *"information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value"* (De Mauro et al. (2016). Coherently, Big Data paradigm indicates technological solutions that allow for real-time processing of large amount of data in motion coming from several sources (e.g. educational environments, social media), through the application of analytics techniques (e.g. classification, clustering) to effectively extract and represent new insights, which may provide a valuable support to the decision making process.

A possible application of Big Data in the online learning domain concerns the evaluation of LS in online courses (Koedinger et al., 2015; Vahdat et al., 2015; Sin and Muthu, 2015), which represents a recent explored research area with limited applications and a potential crucial role in the future of online education (García and Secades, 2013; Noh, 2015).

This paper provides a contribution on this research domain by proposing a novel method and approach to measure the LS in online CL systems. The contribution is based on the analysis of the combination of different data sources such as the answers provided to questionnaires (Sebastiani, 2002) with messages posted within discussion forums (Cope and Kalantzis, 2015) to monitor in real-time and along the entire learning process the level of LS. Moreover, by leveraging Learning Analytics (LA) techniques, the proposed approach allows for enlarging the scope of the current researches on LS that, until now, has been typically oriented to assess the learning experience at the end of a course (Ladyshewsky, 2013).

In such a way, it can be possible to distinguish what has happened in the past from what is going in real time, in the final aim to derive some insights and contribute to design the next future. Thus, the traditional and time-consuming ways to collect feedback through structured questionnaires are complemented by the adoption of non-intrusive techniques that are more preferred (De Montjoye et al., 2013) and do not compromise the reliability of the model (Ortigosa et al., 2013).

In such a conceptual framework, this paper describes a novel method and approach to measure the LS in online CL systems, and presents an innovative archetype named RAMS (RApid Monitoring of learners' Satisfaction) that implements opportunely big data analytics techniques (i.e. sentiment analysis, clustering, and classification) to operationalize the approach.

The proposed contribution may provide universities and corporations with important feedbacks and insights about the level of LS, with the ultimate goal to make more satisfactory their learning experiences (Berenson et al., 2008; Kadry and El Fadl, 2012; Kauffman, 2015; Ma et al., 2016). Indeed, the positive and negative emotions of learners towards an online course (Colace et al., 2014), as well as the satisfaction level of peers (Siemens and Long, 2011), may affect the overall evaluation and can provide mentors and managers with valuable suggestions to make the CL experiences more appealing.

Therefore, the basic research question investigated in this study can be formulated as

follows: *How monitoring and assessing LS in CL environment by adopting a Big Data architecture and learning analytics techniques?*

The paper is structured as follows. Section 2 presents the theory background, section 3 presents the research methodology adopted to develop and test the system, whereas section 4 illustrates the system's architecture with its implementation and test. Finally, section 5 concludes the paper by discussing the main implications, highlighting some limitations, and opening the path for future researches.

2. Theory Background

The background of this article focuses on the conceptual framework of LS in CL domain, integrated by current Big Data approaches and learning analytics techniques supporting them.

2.1 Learning satisfaction in Collaborative Learning

A widely accepted definition of LS refers to the *"perception of approval and accomplishment that learners develop in learning environments"* (Sweeney and Ingram, 2001). It is based on both intrinsic and extrinsic students' motivation (Keller, 1983) and affects learners in both the continuance of the learning process and the demand to conclude the course (Bolliger and Wasilik, 2009).

In technology-enabled environments, LS reflects how positively students perceive their learning experiences, so it is an important indicator of program- and student-related outcomes such as program quality, student retention, and student success in program evaluation (Liao and Hsieh, 2011). Many studies shows that high level of LS can lead to lower drop-out rates, higher persistence, and greater commitment to the program (Kuo et al., 2014).

In CL environments, the frequency of the learning events and the climate of the working groups (Ku et al., 2013) influence the level of LS. Thus, LS can be a proxy of the positive or negative sentiment that participants feel about their CL experiences.

Assessing LS by focusing on key measurable categories is a central issue for online courses (Rivera and Rice, 2002); on this theme, many research contributions have been proposed in literature. Some scholars focused on systemic frameworks, such as Wang (2003) who considered learner interface, learning community, content, and personalization as four key aspects to stimulate LS. Then, Bolliger (2004) identified five key dimensions (communication, technology, course management, course

website, general information), whereas Sun et al. (2008) distinguished six dimensions (learner, instructor, course, technology, design and environment). Besides, Wu et al. (2010) pointed out students' cognitive beliefs, technological environment, and social environment as the main dimensions representing the LS. Finally, Spears (2012) took into account five key dimensions such as technology, physical distance, communications and availability, interaction with instructor and peers, and course design.

Furthermore, other scholars were oriented to analyse the LS from specific views, as Fredericksen et al. (2006) who focused on the type of interaction (with teachers, participation compared to classrooms, interactions with classmates and technical difficulties). Then, So and Brush (2008) examined the relationships of the students' perceived levels of collaborative learning, social presence and overall satisfaction in a blended learning environment. Besides, Paechter et al. (2010) extended LS dimensions also to course design, pace of learning, attendance of the course, and tasks results. Finally, Dejaeger et al. (2012) investigated the construction of comprehensible data mining models and analysed the training related aspects such as the usefulness of the process, the easiness, the efficiency, and the trainer performance. The framework proposed in Table 1 grounds on the above-mentioned researches to provide a holistic view to the phenomenon, which is articulated in five key dimensions.

<Table 1 about here >

To assess the LS, it is required the access to a variety of data sources such as reports, assignments and test results, communication flows and interactions, forums' messages and comments, log files (Bolliger, 2004). These sources represent heterogeneous but potentially valuable data sets that can be analysed through Big Data systems for obtaining in real time useful insights about the level of LS towards each dimension.

2.2 Big Data in educational contexts

Big Data is an umbrella term that encompasses a wide range of concepts including (Gandomi and Haider, 2015): technologies and platforms to address large amount of heterogeneous and fast generated data; systems to store, aggregate and process huge quantity of data; analytical techniques and algorithms to process structured and

unstructured data set (known also as Big Data analytics); new sources of data (social media, mobile devices, sensors) that can be integrated with the knowledge base of the organizations to support decision making; a cultural shift and a new mindset overwhelming business and society.

As initially introduced by Laney (2001), Big Data concept grounds on three main characteristics such as *Volume*, *Velocity* and *Variety*. From 2013, Big Data concept has been enriched by other "V"s (e.g. IBM coined *Veracity*, which represents the unreliability inherent in some sources of data; SAS introduced *Variability*, which refers to the variation in the data flow rates; Oracle introduced *Value* as the benefits/advantages can be obtained by analyzing large volumes of such data). That said, the definition of Big Data adopted in the article relies on the original meaning based on the first three Vs, with a value-oriented perspective. This is the rationale we found in De Mauro et al. (2016), who define Big Data as *"information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value"*.

Since 2012, the development of Big Data paradigm has been very fast for the convergence of four key technological drivers: the increasing of the storage capability, the rapidity of digital processing, the growth of the network speed, and the availability of advanced techniques for data analysis. Moreover, the growing availability and usage of technological systems today allows companies and institutions to experiment new processes to transform vast amount of structured and unstructured data into valuable information (Marr, 2015).

In terms of application contexts, Big Data services are widely adopted in technology, healthcare, and education (Wamba et al., 2015; Siemens and Long, 2011; Soares, 2012). In particular, for the education domain, current LMS and MOOC platforms can potentially produce a huge quantity of data that can be exploited by using data analytics tools, text mining techniques and automated reasoning algorithms (Cope and Kalantzis, 2015) to find new ways to monitor LS.

This would provide valuable insights to design more attractive online courses (Miglietti and Strange, 1998) and ultimately improve the learning experience (Berenson et al., 2008; Kadry and El Fadl, 2012; Kauffman, 2015).

This is usually obtained by relying on Learning Analytics, which are defined as *"the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the*

environments in which it occurs" (Ferguson, 2012). Thus, LA consist in a multidisciplinary approach based on data processing, educational data mining and visualization techniques (Scheffel et al., 2014). LA and Big Data overlap in the methods and techniques used to highlight important and relevant information respect to a subject of interest. In fact, a LA-based approach relies on the implementation of a set of methods and techniques for educational data analysis, including clustering, classification and prediction, process mining, social network analysis, and text mining (Romero and Ventura, 2013). Among these techniques, clustering and classification are the most popular and used (Sin and Muthu, 2015), whereas text mining is increasingly used for the accurate analysis of structured and unstructured data to automatically discover meaningful associations, trends and patterns in large corpora of text (Tane et al., 2004; Fan and Gordon, 2014; Ampofo et al., 2015).

Clustering is often used to identify homogeneous groups of students with similar features and interests to promote collaborative learning (Tang and McCalla, 2002; Obadi et al., 2010). Classification techniques are used to associate groups of students to a known set of categories (Chen et al., 2000) to identify low-motivated learners and find possible actions to reduce drop-outs.

Mainly for unstructured sources of data (e.g. forums, chats, web pages), text mining techniques are used for text categorization, text clustering, document summarization, and sentiment analysis (Tane et al., 2004). Among these, since affective and emotional factors may influence learners' satisfaction (Shen et al., 2012), sentiment analysis is particularly important to rate online courses and define possible strategies to improve them (Bharathisindhu and Brunda, 2014). The most diffused methods adopted for sentiment analysis are machine-learning methods (supervised or unsupervised) and lexicon-based methods (Medhat et al., 2014). The main focus of these techniques is the extraction of learners' opinions and sentiment expressed within blogs and forums where services and contents are evaluated (Kechaou et al., 2011; Wen et al., 2014) to identify the most critical issues (Zarra et al., 2016) and design coherently specific actions and resources (Ortigosa et al., 2014).

2.3 Big Data Analytics for learner's satisfaction

Though the research on the application of Big Data Analytics to the CL domain is active and evolves rapidly, it has still limited applications on learners' satisfaction, which represents an ongoing issue that can open promising impacts at research and

practitioner levels. Nonetheless, some interesting researches are related to the application of Learning Analytics techniques, such as the study conducted by Thomas and Galambos (2004) that identifies which characteristics and experiences affect learners' satisfaction using decision trees and regression analysis to process the opinions collected among the students. Later, Chen and Chen (2009) developed a tool based on six computational intelligence theories (i.e. correlation analysis, fuzzy clustering, grey relational analysis, k-means clustering, fuzzy association rule mining, fuzzy inference) to help teachers to assess the individual learning satisfaction by using only the online learning portfolios. Dejaeger et al. (2012) carried out a research focused on identifying the learners' satisfaction factors by applying Support Vector Machines (SVM) algorithm, neural networks, decision trees and regression analysis, and build an overall model to identify the main drivers of students' satisfaction and support the decision-making process. Sentiment analysis has been used to process Facebook pages to determine the learners' mood in a non-intrusive way, in the aim to enrich the user models for adaptive learning systems (Martín et al., 2012). Recently, Lustigova and Novotna (2016) applied text mining suites on information related to learners' satisfaction, and discovered that some of the limitations encountered (e.g. the usage of multilingual and heterogeneous sources) may be effectively faced by adopting proper Big Data architecture (beyond just techniques and algorithms).

A typical use of Learning Analytics concerns the support to design personalized learning approaches (Buckingham Shum, 2014; Knight et al., 2014). Survey-based analytics are used to establish students' learning dispositions and provide personalised and timely feedback on their readiness to undertake specific courses (Deakin Crick et al., 2013). Ozpolat and Akar (2009) proposed an automatic student modelling method based on a keyword mapping and clustering. Along the same line, Lee (2012) presented a learning preference forecasting model to classify learners' profiles by preference degrees. Finally, Hsu (2008) applied clustering and association rule algorithms to provide tailored contents in English courses.

Although these researches have experienced the application of Learning Analytics techniques, at our knowledge no study ensures the real-time dimension in the LS assessment process, as well as leverages Big Data paradigm.

In such a perspective, this work aims at providing a novel contribution by proposing a method and approach that leverage Big Data characteristics to assess LS in collaborative online courses by using Learning Analytics techniques, and Big Data

architecture and infrastructure.

3. Research Methodology

The research methodology adopted in this study is based on the design science research, which plays a central role in the development and management of information systems (van Aken et al., 2016; Hevner and Chatterjee, 2010; Hevner et al., 2004). Coherently with the ground of design science research, which aims at implementing processes or systems dealing with problems of a significant practical relevance, the activities carried out and the research methodology adopted rely on the three constituting cycles (Peffers et al., 2006; Hevner, 2007): the relevance cycle (focused on the identification of requirements, evaluation criteria, and field testing); the design cycle (focused on building and evaluating the design artefacts and processes); and the rigor cycle (focused on providing the theoretical grounds of experiences and artefacts that define the state-of-the-art in the application domain of the research).

As for the *relevance cycle*, the identification of requirements relies on the importance of the assessment of LS in online courses, together with the awareness of the big potential related to the application of the Big Data paradigm to monitor this issue in real time, and define an appropriate intervention strategy. Coherently with this environment, a simple scenario inspired to the Big Data paradigm has been envisioned, where it would be possible to design a Big Data architecture, implement it by using open source components, and perform a controlled field test by involving a small sample of online learners to evaluate the overall system and the innovative functionalities based on LA techniques. The phase of controlled experiment was focused on the evaluation of the results obtained from sentiment analysis, classification algorithms and clustering techniques applied to the forum messages and questionnaires.

As for the *rigor cycle*, the theoretical background of the experience and tools that define the state-of-the-art in the application domain highlights the limited adoption of Big Data applications (in terms of both architecture, infrastructure, and techniques) for the on-going evaluation of the LS within CL experience. Indeed, previous research contributions refer mainly to the application of Learning Analytics techniques, without implementing a complete Big Data architecture and infrastructure for monitoring LS of online courses. This brought to define the research goal that consists

in the investigation of new methods and approaches for a real-time assessment of the level of LS of online course by leveraging Big Data architecture, infrastructure and techniques. Moreover, the dimensions affecting the LS in online courses have been identified through a specific literature review, which allowed defining the LS model used in the controlled experiment. The results achieved through the field test of the system contribute to the ongoing debate about the application of the Big Data paradigm in the online education domain.

As for the *design cycle*, the development of the artefact included the selection and integration of open source software components for building the Big Data infrastructure, as well as the implementation of ad-hoc customization and software plug-ins to collect and process data, and visualize the results. The demonstration of the artefact has been done by testing the functionalities of the integrated system to check both the data management and data visualization services. At this purpose, a controlled experiment (Hevner et al., 2011) has been defined by collecting a reduced data set composed by interactions and messages of ten students attending an online course on crowdfunding for a three-week period. This data set has been processed just to test on the field the system developed and show its functioning, without any presumption to provide an example of a Big Data context or validate the tool, for which it is necessary the involvement of a huge number of students that generate high volume of data. Finally, the evaluation of the archetype developed has been done through a mixed two-step approach. The first step relies on a two-day check between what the system revealed on the sentiment about the course and what learners expressed in face-to-face interactions with the tutor. The second step relies on periodic face-to-face checks made by the tutor in correspondence of the system alerts. By preserving the final goal of the design science research that is the generation of artefacts that are pragmatically valid, with a robust research basis and a rigorous method for their design and evaluation, and a significant relevance for researchers and practitioners (van Aken et al., 2016), the design cycle adopted in this study relies on just one round of controlled experimentation and field test, thus missing the opportunity of further steps for a possible generation-design-test of alternatives.

4. Findings: a Big Data application for RApid Monitoring of learners Satisfaction (RAMS)

RAMS system, which stands for RApid Monitoring of learners' Satisfaction, and its

application are the main results presented in this section. RAMS is a tool allowing for automatic, continuous and real-time monitoring of LS. It leverages a Big Data infrastructure and relies on Learning Analytics approach for the effective management and the real-time processing of large amount of data, coming from several sources and having different formats. It adopts data mining techniques (i.e. clustering, classification, and text mining), to provide mentors and learning managers with valuable insights about the level of LS during the delivery of an online course. In order to demonstrate the functioning of the implemented services, RAMS has been tested within a controlled experiment (Hevner et al., 2011) involving ten students enrolled into a three-week online course. The test has been based on an open source LMS as the main data provider. The following sections show the RAMS architecture with a focus on the Big Data infrastructure and LA techniques adopted, as well as the results related to the system implementation and test.

4.1 System architecture

RAMS architecture is composed by three components (i.e. RAMS Server, Message Bus, and LMS RAMS plugin) shown in Figure 1 and described below, which implement both Data Management and Data Analytics processes (Labrinidis and Jagadish, 2012). Data Management refers to processes and technologies to acquire, store, prepare and retrieve data, whereas Data Analytics refers to techniques for the analysis of data and the extraction of knowledge and insights.

<Figure 1 about here >

The *RAMS server* is built on a Big Data infrastructure and processes textual information units by using text analytics techniques. The core element is the *RAMS Analysis Engine*, which includes: a *classification engine* that uses supervised text mining algorithms to classify the information units according to LS dimensions; a *sentiment analysis engine* that calculates the positive or negative sentiment for each information unit; and a *cluster engine* that extracts the hot topics identified within the set of information units.

Beyond the RAMS Analysis Engine, RAMS server includes also a *Training Data Set Repository* used to train supervised learning algorithms (i.e. sentiment and

classification). The knowledge and results extracted from the information units processing are stored in the *Data Repository*.

Finally, the *Visual Analytics component* uses info-graphics to provide mentors and learning managers with synthetic information related to the level of LS.

The *Message Bus* is in charge of streaming data to the RAMS server. It is made up by two elements: a *Streaming Event Data* component that guarantees the real-time streaming of data through a reliable and scalable service that captures as "event" any new data produced by the LMS; and a *Message Broker*, which is a scalable and durable publish-subscribe messaging system that stores into an asynchronous queue the events received by the Streaming Event Data component, and sends them to the RAMS server for their processing.

Finally, the *LMS RAMS plugin* is in charge of visualizing (within the LMS) the results generated by the Visual Analytics component.

These three components process two main data sources, which are included into the LMS Repository: the course's discussion forums where learners interact with teachers, mentors and peers; and the learners' feedback gathered through an open-answer questionnaire that collects information about the five identified dimensions of the LS. Both data sources are monitored by the Streaming Event Data component in order to extract the information units and allow the application of Learning Analytics techniques to each LMS data source.

Table 2 provides a synthetic view about the specific LA techniques that are applied to the different data sources.

<Table 2 about here >

4.2 System implementation

The RAMS implementation has been realized by using open source components, such as:

- *Moodle* (www.moodle.org) as LMS used for extracting data and hosting the RAMS plug-in;
- *Apache Flume* (http://flume.apache.org) as streaming data flows component for collecting and aggregating data generated within the different LMS data sources;

- *Apache Kafka* (http://kafka.apache.org) as message broker for implementing a publish-subscribe messaging system to communicate with the RAMS Analysis Engine;
- *RAMS Analysis Engine* that is the component developed with Spring Boot and RapidMiner (http://rapidminer.com) for implementing big data analytics processes (sentiment analysis, clustering, and classification);
- *ElasticSearch* (http://www.elastic.co) as scalable data repository and full text search engine;
- *Kibana* (www.elastic.co/products/kibana) as visual analytics tool for real-time data visualization;
- *RAMS Moodle plugin* as ad-hoc component developed and embedded within the LMS for the interactive visualization of results.

The design choices of the RAMS architecture allow for establishing a multi-tenant system to be offered as SaaS (Software as a Service) for a real-time processing of large amount of data coming from different LMS systems. To achieve this goal, the system's architecture uses highly scalable and reliable software components. More in detail, RAMS server uses Apache Flume to read forum's messages and questionnaires through a Flume "database source", and then streams them to Kafka by using a configured Flume sink. Kafka manages messages produced by the sink by using two topics: the first one enqueues messages related to the discussion forum for sentiment analysis, clustering and classification; whereas, the second one enqueues the messages related to the questionnaire for sentiment analysis and clustering. The use of Apache Flume and Apache Kafka for the data ingestion phase allows RAMS for moving large amounts of data from many different sources in a reliable and high-availability manner, as well as to stream data flow with robustness and fault-tolerance with high reliability mechanisms. It ensures also the decoupling between data production (in charge to the LMS) and data consumption process (in charge to the RAMS Analysis Engine).

As for the visualization, the results of the data analytics processing are stored in ElasticSearch indexes in order to provide Kibana with a fast and reliable access to this information, and generate info-graphics.

Coherently with the logic of Big Data paradigm, as soon as data are created and stored in the LMS, they are immediately processed by RAMS Server and the results are presented through info-graphics. The RAMS Plugin, which is implemented in Moodle, embeds

Kibana dashboards directly in the LMS to show the results in real-time.

Figure 2 shows data flows among the open source components used to build the system.

<Figure 2 about here >

By focusing on the RAMS Analysis Engine, it adopts filtering operators to clean the data set (e.g. HTML tags removal, lower case letters transformation, tokenization, stop words removal, dictionary-based stemming, etc.) before to extract the text features and process the data. Then, as for the text features extraction, RAMS analysis engine uses TF-IDF defined as (Na et al., 2004):

$$TF * Log\left(\frac{N}{DF}\right)$$

where TF is the number of times the term occurs in the current processed document, N is the number of term's occurrences in the whole dataset, and DF is the document frequency, that is the number of documents in the dataset containing the term.

Afterwards, the text features extracted are used for implementing the analytics techniques such as sentiment analysis, clustering, and classification.

As for the sentiment analysis, the system uses machine-learning algorithms (i.e. linear SVM and Cross Validation) to classify textual information streams in the aim to define in real time the learner's sentiment about the course. The text mining process is supervised, so that the model has been previously created by using a 15,000 balanced textual source.

For what concerns the text classification, it allows the automatic labelling of textual data sources with predefined categories based on their content. Coherently with the approach used for the sentiment analysis, text classification has been implemented by using a statistical supervised learning algorithm (i.e. SVM). Moreover, since the five dimensions characterizing the LS model represent the target labels, the text classification implemented by the RAMS Analysis Engine is a multi-label classification problem. To face this kind of problem by using SVM, the "one-against-one" method has been adopted with a "max wins" voting strategy, since it reveals more suitable for practical use (Hsu and Lin, 2002).

The choice to use SVM as core algorithm for the Analysis Engine is supported by its great performances in binary classification problems when compared with kNN (k-Nearest

Neighbour), LSSF (Linear Least Square Fit), Neural Networks (Yang and Liu, 1999; Lewis et al., 2004), although the training and optimization phase can be expensive from the computational point of view.

Finally, as for the clustering, it relies on kMeans algorithm to group the feedbacks that learners provide by filling each section of the questionnaire, in the final aim to highlight the "hot-topics". kMeans algorithm is an unsupervised clustering technique based on the partitioning of data into 'k' subsets, where each data element is assigned to the closest cluster based on the distance of the element from the centre of the cluster. In order to use k-means clustering with textual data, a text-to-numeric transformation has been required, implemented through the TF-IDF algorithm. The KMeans clustering adopted extracts three clusters of terms for each section of the questionnaire, and three clusters of terms for each set of labelled posts in the forum.

4.3 System test and evaluation

The test of the whole system aims at showing the functioning of the implemented services to evaluate the level of LS, without any presumption to provide an example of a real Big Data context or to validate the tool, for which it is necessary the involvement of a huge number of students that generate high volume of data.

Indeed, the system has been tested through a controlled experiment (Hevner et al., 2011) based on a reduced data set and involving ten students attending an online course on crowdfunding for a three-week period.

The test started with the population of the LMS data sources associated to the forum and the final questionnaire, in order to check properly data ingestion and messages flows. A discussion forum was thus created in Moodle, as well as a questionnaire to collect the opinions that learners express once they finish the course. Table 3 illustrates the content of the questionnaire centred on the key dimensions identified for the LS.

<Table 3 about here >

During the controlled experiment, as soon as learners posted the messages in the discussion forum, the RAMS Server extracted them to calculate sentiment, define clusters, and make classification. Results of data processing were updated in real-time, were presented by using infographics and were accessible in real-time to teachers,

tutors and learning manager through the RAMS plug-in.

RAMS dashboards related to the analysis of forum's messages are reported in the following figures. Specifically, Figure 3 shows the forum sentiment dashboard including four sections: the total sentiment rate of the entire forum (3a); the details respectively about the sentiment rate per learner (3b) and per LS topic (3c); the details about the filter of the sentiment rate per learner and per LS topic (3d).

<Figure 3 about here >

The four sections of the dashboard may provide insights about the overall judgement of the course (3a), the most "critical" learners (3b), the most criticized topic according to the LS model (3c), and the individual learners' opinion about each topic of the LS model (3d). Indeed, it shows infographics that may support answering to the following questions: how many learners do have critical opinions? Who are they? Which specific topic of the LS model do they criticize the most? Which is actually the most critical dimension for LS?

The dashboard allows also interactive drill-down operations to select data and visualize the results by focusing on a specific learner. As an example, in case of few active learners, data are shown by using a pie chart (figures 3b and 3d), while in case of a typical Big Data scenario involving hundreds of users, the system will use tables to visualize the list of users with associated sentiment. Afterwards, with a drill down operation, the system can show data related to a selected learner, as illustrated in Figure 4.

<Figure 4 about here >

Figure 5 shows the evolution of the sentiment rate of the forum's messages along time, and specifically per day (5a) and per week (5b). These insights can be useful to highlight critical periods where the negative sentiment is predominant respect to the positive one. Figure 5 shows also the daily sentiment rate per each topic of the LS model (5c), in order to understand which aspects of the course are generating negative feelings from the learners' point of view.

<Figure 5 about here >

The analysis of the open answer questionnaires, filled by the learners at the end of the course, provided insights on the total sentiment rate of the questionnaire (Figure 6a), the distribution of positive and negative opinions per each section (Figure 6b), and the result of the clustering applied to all the five sections. This dashboard provided insights about the sentiment expressed by learners at the end of the course, as well as on the hot topics that emerged from the questionnaires.

<Figure 6 about here >

Finally, Figure 7 shows the results of clustering applied to each section of the questionnaires, useful to understand which specific hot topics emerged in each section of the questionnaire.

<Figure 7 about here >

The system accuracy has been evaluated through a mixed approach, which integrates periodic observations made by tutor with ad-hoc checks in correspondence of system alerts. Actually, during the three-week course period, the tutor monitored face-to-face each learner every two days in order to reveal both the individual and general sentiment about the course. Afterwards, these data have been compared with the results shown in the RAMS dashboards and related to the same periods. The result of the comparison reveals the correspondence between what the tutor registered face-to-face and what the system visualized online.

Moreover, for what concerned the ad-hoc checks made in correspondence of the system alerts, they are referred to the two critical periods shown in figure 8 in which the negative sentiment represented with a dark-grey bar is prevalent respect to the positive one with a light-grey bar. These periods corresponded to the course start-up and to the launch of the project work. After a face-to-face check made by the tutor at these events, also in this case the results generated by the system were corresponding to what the tutor revealed. Indeed, the two critical events generated some misunderstandings between students related to some online activities, and the immediate observation of their sentiment allowed teacher to intervene immediately to solve the problems and restore a positive climate in the learning environment.

<Figure 8 about here >

Although the system test has been conducted in a controlled setting, with a small number of students involved and not a huge volume of data generated, the test showed that:

- RAMS services offer the expected functionalities, and the analytics processes implemented guarantee the correct estimation of the learners' sentiment;
- RAMS supports correctly the LS assessment process in real-time;
- RAMS can support the decision making process of tutors, teachers and learning managers during a course.

5. Discussions and conclusion

This paper proposes a novel approach to assess in real time the LS within online courses, by leveraging Big Data infrastructure and LA techniques. The studies cited in the literature background show that monitoring LS in online courses is a key activity for successful CL experiences (Gunawardena et al., 2001; Wu et al., 2010). These studies highlight also that, although the research on the application of Big Data techniques to the CL domain is very active and evolves rapidly, the focus on LS is still limited and, at our knowledge, no study fully implements Big Data architecture, infrastructure and functionalities for monitoring LS in online courses.

In such a view, this paper also presents and discusses the design, implementation and test of an innovative system called RAMS, which leverages fully the Big Data paradigm for real-time and continuous monitoring of LS within online courses. The system design covers the four analytical steps that characterize a typical Big Data application (Flyverbom and Madsen, 2015), which are: data production (realized directly by the students through a forum and a questionnaire); data structuring (through the proper design of data repository where the contents are stored); data distribution (by feeding the analysis engine that elaborates sentiment, clusters and topics); and data visualization (realized by interactive and personalized dashboards that show the results).

From a Learning Analytics perspective, the proposed system further explores this field by contributing to bridge the gap between efficient data acquisition, immediate analysis to provide real time feedbacks (Long and Siemens, 2011), and transformation

of data into meaningful information through contextual analysis (Elias, 2011). These are crucial challenges for providing teachers and managers with effective means to understand and optimize the context in which the overall learning processes occur (Ferguson, 2012). Indeed, RAMS implements techniques for data collection and gathering, real-time data processing and visualization, which are typical characteristics of Learning Analytics (Nunn et al., 2016). Finally, up to our knowledge, the system presented is pioneering the integration of Learning Analytics techniques with a full Big Data architecture and infrastructure for monitoring LS within online courses.

In fact, RAMS architecture has been designed and implemented to easily connect simultaneously numerous different data sources, and so being ready to ingest huge volume of data. Data sources can be traditional LMS (e.g. Moodle) or MOOC platforms; once data are generated within these systems, they are immediately streamed so ensuring velocity and real-time processing. At this purpose, Flume and Kafka offers sub-second-latency event processing: Flume is one of the most widely adopted data processing frameworks for real-time data analytics, whereas Kafka is a low-latency platform for real-time processing that is largely used to process streaming data within complex infrastructures. Moreover, by configuring properly the connection to the different data sources, Flume ensures management of data of many forms (e.g. structured data coming from a forum or questionnaire created in Moodle as implemented in RAMS, as well as unstructured data retrieved from social network or text scraped from the web). Definitely, the choices adopted for designing and implementing RAMS are aligned to the Big Data paradigm, and in particular with the three founding Vs (Volume, Velocity, and Variety).

As for the test, the proposed system has been tested through a controlled experiment (Hevner et al., 2011) involving few online students and a small dataset, aimed at showing the RAMS functionalities and their embedded value, without any presumption to provide an example of a Big Data context, for which it is necessary the involvement of a huge number of students that generate high volume of data. Indeed, the test showed that the system may support teachers and tutors to monitor in real time the LS and know the learners' sentiment, in the aim to promptly put in place possible strategies and recovering actions to solve critical situations or plan the next future. Actually, they can identify those learners who explicitly manifest negative opinions and can provide them with punctual feedbacks tailored to their problems, thus supporting continuously the progress of their learning process. Furthermore, the

analysis of the topics expressed by those learners who completed the course and answered to the final questionnaire represent valuable insights to better support the other learners who are still involved in the same course, but also to improve the LS dimensions and make more attractive the entire course.

Definitely, the proposed approach and system, by providing reliable and real time information on both single learners and overall learning context, contribute to enhance the quality of decision making process (McAfee and Brynjolfsson, 2012) that evolve from being based on perceptions or "HiPPO" effect (high-paid-person-opinions) to be solidly driven by real data, objective measures, and instantaneous processing. This represents a valuable support to improve significantly the educational processes, and explore new paths to follow in the new educational era (García and Secades, 2013).

However, this paper has also some limits. The involvement of few users during the system test represents a first limit of this research in terms of generalization of the results. Besides, the execution of just one cycle of field test, which was mainly due to both the absence of further online courses for the same students and the absence of further students for the same course, hinders a robust validation, optimization and generalization of the results.

Further research directions arising from this article concern the discovering of possible relationships between the polarity of the "sentiment" of a course with the final "mark" of learners, in order to test the predictability of the performance of the learning experience. Besides, the integration of the internal sources of learners' data (e.g. forum messages, login information, enrolled courses) with the external sources (e.g. contents posted in social media) may provide a multidimensional view of learner's profile that is of potential interest for both learning providers and people managers. Indeed, the application of text mining techniques and algorithms for mining, correlating and visualizing digital traces from disparate data sources can enable the measurement, prediction and governance of complex phenomena, including learning processes (Boyd and Crawford, 2012).

Another research direction more focused on technological issues concerns the adoption of deep learning techniques for classification purposes, by implementing Recursive Neural Tensor Networks for the sentiment estimation (Socher et al., 2013).

Finally, in terms of business impact, further analysis could be devoted to transform RAMS into a Software-as-a-Service platform by implementing multi-tenant features, thus enabling distinct institutions and organizations to access to personal and

customizable dashboards that show the level of LS of their own courses.

References

Abdrabo, M., Elmogy, M., Eltaweel, G., & Barakat, S. (2016). Enhancing Big Data Value Using Knowledge Discovery Techniques. *International Journal of Information Technology and Computer Science*, 8, 1-12

Aitken, N.D. (1982). College student performance, satisfaction and retention: Specification and estimation of a structural model. *The Journal of Higher Education*, 53(1), 32- 50.

Ampofo, L., Collister, S., O'Loughlin, B. & Chadwick, A. (2015). Text Mining and Social Media: When Quantitative Meets Qualitative and Software Meets People. *Innovations in Digital Research Methods*, 161-91.

Barkley, E., Cross, P.K., & Major, C.H. (2004) *Collaborative learning techniques: a handbook for college faculty*. Jossey-Bass, San Francisco, CA.

Berenson, R., Boyles, G., & Weaver, A. (2008). Emotional intelligence as a predictor for success in online learning. *International Review of Research in Open & Distance Learning*, 9(2), 1–16

Bharathisindhu, P. & Brunda, S. S. (2014). Identifying E-Learner's Opinion Using Automated Sentiment Analysis in E-Learning. *IJRET: International Journal of Research in Engineering and Technology*, 3(1), 2319-2322.

Bolliger, D. U. & Wasilik, O. (2009). Factors influencing faculty satisfaction with online teaching and learning in higher education. *Distance Education*, 30(1), 103-116

Bolliger, D.U. (2004). Key Factors for Determining Student Satisfaction in Online Courses. *International Journal on E-Learning*, 3(1), 61-67.

Boyd, D. & Crawford, K. (2012) Critical Questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15, 662-679

Buckingham Shum, S. (2014). Personalisation: A learning analytics lens. Paper presented at the Universitas 21 (U21) Educational Innovation Conference, Sydney, Australia

Bughin, J. (2016). Big data, Big bang? *Journal of Big Data*, 3(1), 1-14

Chen, C.M. & Chen, M.C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education*, 52(1), 256-273

Chen, G., Liu, C., Ou, K., & Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. *Journal of Educational Computing Research*, 23(3), 305-332

Colace, F., De Santo, M. & Greco, L. (2014). SAFE: A Sentiment Analysis Framework for E- Learning. *International Journal of Emerging Technologies in Learning*, 9(6), 75-91.

Conole, Grainne, Alevizou, & Panagiota (2010). *A literature review of the use of Web 2.0 tools in Higher Education*. HEA Academy: York, UK.

Cope, B., & Kalantzis, M. (2015). Sources of Evidence-of-Learning: Learning and assessment in the era of big data. *Open Review of Educational Research*, *2*(1), 194-217.

De Mauro, A., Greco, M. & Grimaldi, M. (2016) A formal definition of big data based on its essential features, *Library Review*, 65(3), 122-135.

De Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting people personality using novel mobile phone based metrics. Social Computing, Behavioral– Cultural Modeling and Prediction. *Lecture Notes in Computer Science*, 7812, 48–55.

Deakin Crick, R., Goldspink, C., & Foster, M. (2013). Telling identities: Learning as script or design? Learning emergency discussion paper.

Dejaeger, K., Goethals, F., Giangreco, A., Mola, L. & Baesens, B. (2012). Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, 218(2), 548–562.

Donohue, T.L. & Wong, E.H. (1997) Achievement motivation and college satisfaction in traditional and non-traditional students. *Education*, 118(2), 237-243

Elia, G., & Poce, A. (2010). *Open Networked i-Learning*, Berlin: Sperling.

Elia, G., Margherita, A., & Taurino, C. (2009) Enhancing Managerial Competencies through a Wiki-Learning Space. *International Journal of Continuing*

Engineering Education and Life-Long Learning, 19(2/3), 166-178.

Elias, T. (2011). Learning analytics. Learning.

Fan, W. and Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, 57(6), 74-81.

Ferguson, R., Learning analytics: Drivers, developments and challenges, Int. J. Technol. Enhan. Learn. 4 (2012), 304–317.

Flyverbom, M., & Madsen, A.K. (2015) Sorting data out: Unpacking big data value chains and algorithmic knowledge production. *Die Gesellschaft der Daten. Über die digitale Transformation der sozialen Ordnung*

Frederiksen, E., Pickett, A., & Shea, P. (2006). Student satisfaction and perceived learning with online courses: Principles and examples from the SUNY learning network. *Journal of Asynchronous Learning Networks,* 4(2), 2-31.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, 137-144

García, O.A., & Secades, V.A. (2013). Big data and learning analytics: A potential way to optimize elearning technological tools. *In Proceedings of the International Conference e- Learning 2013, Part of the IADIS Multi Conference on Computer Science and Information Systems 2013*, MCCSIS 2013, 313-317.

Gunawardena, C. N., Nolla, A. C., Wilson, P. L., López-Islas, J. R., Ramírez-Angel, N. & Megchun-Alpízar, R. M. (2001). A cross-cultural study of group process and development in online conferences. *Distance Education (An International Journal),* 22(1), 85-121.

Hevner, A., & Chatterjee, S. (2010). Design science research in information systems. In *Design Research in Information Systems*, 9-22. Springer US.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75–105.

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415-425

Hsu, M. (2008) A personalized English learning recommender system for ESL students. *Expert Systems with Applications*, 34 (1), 683-688

Kadry, M. A., & El Fadl, A. R. M. (2012). A proposed model for assessment of social networking supported learning and its influence on learner behaviour. *In Proceedings of Interactive Mobile and Computer Aided Learning (IMCL)*, 2012 International Conference, 101-108.

Kauffman, H. (2015). A review of predictive factors of student success in and satisfaction with online learning. *Research in Learning Technology*, 23.

Kechaou, Z., Ammar, M. B. & Alimi, A. M. (2011). Improving e-learning with sentiment analysis of users' opinions. *In Proceedings of 2011 IEEE Global Engineering Education Conference (EDUCON)*, 1032-1038, IEEE.

Keller J. (1983). Motivational design of instruction. In C. Reigeluth (Ed.), *Instructional design theories and models: An overview of their current status*, 386-434. Hillsdale, NJ: Erlbaum.

Kirschner, P. A., Kreijns, K., Phielix, C., & Fransen, J. (2015). Awareness of cognitive and social behaviour in a CSCL environment. *Journal of Computer Assisted Learning*, 31(1), 59-77.

Knight, S., Buckingham Shum, S., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics*, 1(2), 23-47.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *WIREs Cognitive Science*, 6, 333–353.

Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.). *Handbook of Educational Data Mining* (43-56). CRC Press..

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, 19(3), 335-353.

Ku, H., Tseng, H.W., & Akarasriworn, C. (2013).Collaboration factors, teamwork satisfaction, and student attitudes toward online collaborative learning. *Computers in Human Behavior*, 29, 922–92.

Kuo, Y.C., Walker, A.E., Schroder, K.E., Belland, B.R. (2014). Interaction, Internet self-efficacy, and self-regulated learning as predictors of student satisfaction in

online education courses. *The Internet and Higher Education*, 20, 35-50

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12), 2032-2033.

Ladyshewsky, R. (2013). Instructor presence in online courses and student satisfaction. *The International Journal for the Scholarship of Teaching and Learning*, 7(1), 1-23.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).

Lee, Y. (2012) Developing an efficient computational method that estimates the ability of students in a web-based learning environment. *Computers & Education*, 58 (1), 579–589

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(Apr), 361-397.

Liao, P. W., & Hsieh, J. Y. (2011). What influences Internet-based learning?. *Social Behavior and Personality: An International Journal*, 39(7), 887-896.

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 31–40.

Lustigova, Z., & Novotna V. (2016). Advantages and Limits of Text Mining Software for Analysis of Students' Satisfaction in Online Education. *Conference Proceedings. The Future of Education*, 1(2)

Ma, B., Zhang, N., Liu, G., Li, L., & Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing & Management*, 52(3), 430-445.

Marr, B. (2015) *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Wiley.

Martín, J. M., Ortigosa, A., & Carro, R. M. (2012) SentBuk: Sentiment analysis for e-learning environments, *Proceedings of International Symposium on Computers in Education (SIIE),* 29-31.

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), 61-67

Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.

Miglietti, C.L. & Strange, C.C. (1998) Learning styles, Classroom environment preferences, teaching styles, and remedial course outcomes for underprepared adults at a two-year college. *Community College Review*, 26(1), 1-19

Na, J.C., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004) Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *Proceedings of the Conference of the International Society for Knowledge Organization (ISKO)*, 49-54. Wurzburg, Germany.

Noh, K. S. (2015). Plan for Vitalisation of Application of Big Data for e-Learning in South Korea. *Indian Journal of Science and Technology*, 8(S5), 149–155.

Nunn, S., Avella, J. T., Kanai, T., & Kebritchi, M. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. Online Learning, 20(2).

Obadi, G., Drázdilová, P., Martinovic, J., Slaninová, K., & Snásel, V. (2010) Using spectral clustering for finding students' patterns of behavior in social networks, *Proceedings of International Workshop on Databases, Texts, Specifications, and Objects (DATESO)*, 567, 118-130.

Ortigosa, A., Carro, R, & Quiroga, J. (2013). Predicting user personality by mining social interactions in Facebook. *Journal of Computer and Systems Sciences.* Special issue on Intelligent Data Analysis.

Ortigosa, A., Martín, J. M. & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31, 527-541.

Ozpolat, E. & Akar, G. (2009) Automatic detection of learning styles for an e-learning system. *Computers & Education*, 53 (2), 355–367

Paechter, M., Maier, B., & Macher, D. (2010). Students' Expectations Of, And Experiences In Elearning: Their Relation To Learning Achievements And Course Satisfaction. *Computers & Education*, 54, 222-229.

Pearson, T. & Wegener, R. (2013). *Big Data: The organizational challenge*. www.bain.com/Images/BAIN_BRIEF_Big_Data_The_organizational_challenge.pdf

Peffers, K., Tuunanen, T., Gengler C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The design science research process: A model for producing and presenting information systems research. In *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006)*, 83-106

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45-77.

Ravenscroft (2011). Dialogue and Connectivism: A New Approach to Understanding and Promoting Dialogue-Rich Networked Learning. *The International Review of Research in Open and Distributed Learning*, 12 (3)

Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004). Integrating personalization in e-learning communities. *International Journal of Distance Education Technologies* (IJDET), 2(3), 47-58

Rivera, J. & Rice, M. (2002). A comparison of student ourcome and satisfaction between traditional and web based course offering. *Online journal of Distance Learning Administration*, 5 (3).

Romero, C. and Ventura, S. (2013). Data Mining in Education. *WIREs Data Mining and Knowledge Discovery*, 3, 12-27

Scheffel, M., Drachsler, H., Stoyanov S., & Specht, M. (2014). Quality indicators for learning analytics. *Educational Technology & Society*, 17(4), 117−132.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

Shen, L., Wang, M., & Shen, R. (2012). Affective e-learning: Using ‘‘Emotional’’ data to improve learning in pervasive learning environment. *Educational Technology & Society*, 2, 176–189.

Siemens, G. (2004). Connectivism: A Learning Theory for the Digital Age. *International Journal of Instructional Technology & Distance Learning*, 2(1)

Siemens, G. & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE review*, 46(5), 30.

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics - A literature review. *ICTACT Journal on Soft Computing*, 5(4), 1-35.

So, H. J., & Brush, T. A. (2008). Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. *Computers & Education*, 51, 318-336.

Soares, L. (2012) The rise of big data. *EDUCAUSE Review*, 47(3), 60.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 1631-1642.

Spady, W. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1), 64-85.

Spears, L. R. (2012). Social Presence, Social Interaction, Collaborative Learning, and Satisfaction in Online and Face-to-Face Courses. *Graduate Theses and Dissertations.* Paper 12976.

Su, A. Y., Yang, S. J., Hwang, W. Y., & Zhang, J. (2010). A Web 2.0-based collaborative annotation system for enhancing knowledge sharing in collaborative learning environments. *Computers & Education*, 55(2), 752-766.

Sun, P., Tsai, R. J., Finger, G., Chen, Y, & Yeh, D. (2008). What drives a successful e-learning? An empirical Investigation of the critical factors influencing learning satisfaction, *Computer & Education,* 50, 1183-1202.

Sweeney, J.C., & Ingram. D. (2001). A Comparison of Traditional and Web-Based Tutorials in Marketing Education: An Exploratory Study. *Journal of Marketing Education*, 23(1), 55-62.

Tane, J., Schmitz, C., & Stumme, G. (2004) Semantic resource management for the web: an e-learning application. *Proceedings of International Conference of the WWW*, 1–10. New York.

Tang, T.Y., & McCalla, G. (2002). Student Modeling for a Web-based Learning Environment: a Data Mining Approach. *Proceedings of 18th National Conference on Artificial Intelligence*, 967-968. American Association for Artificial Intelligence, Menlo Park.

Thomas, E. H., & Galambos, N. (2004). What Satisfies Students?: Mining Student-Opinion Data with Regression and Decision Tree Analysis. *Research In Higher Education*, 45(3), 251-269.

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). Advances in learning analytics and educational data mining. In *Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges (Belgium), 297-306.

van Aken, J., Chandrasekaran, A., & Halman, J. (2016). Conducting and publishing design science research: Inaugural essay of the design science department of the Journal of Operations Management. *Journal of Operations Management*, 47,1-8

Vygotsky, L. (1978). *Interaction between Learning and Development In Mind in Society,* Harvard University Press, Cambridge, MA.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G. & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234-246

Wang, Y. S. (2003). Assessment of learner satisfaction with asynchronous electronic learning systems. *Information & Management*, 41(1), 75-86.

Wen, M., Yang, D. & Rosé, C. P. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us. In *Proceedings of Educational Data Mining 2014*.

Wright, J. (2015). Personalised learning: an overview.

Wu, H., Tennyson, R.D., & Hsia, T. (2010). A Study of Student Satisfaction in a Blended E- Learning System Environment. *Computers and Education*, 55, 155-164

Yang, Y., & Xin, Liu (1999) A re-examination of text categorization methods. *Proceedings of the 22nd International Conference ACM SIGIR on Research and Development in Information Retrieval.*

Zarra, T., Chiheb, R., Faizi, R., & El Afia, A. (2016). Using Textual Similarity and Sentiment Analysis in Discussions Forums to Enhance Learning. *International Journal of Software Engineering and Its Applications*, 10(1), 191-200.

Table 1. Key dimensions influencing Learner Satisfaction

Dimensions	**Description**	**Reference Authors**
Course structure and contents	- Course quality (clarity of contents and learning goals, up-to-date level and usefulness of contents) - Course flexibility - Instructor attitude toward e-learning	Wang, 2003; Sun et al., 2008.
Technology and Support	- Usability of the platform - Usefulness of the services - Easiness to access - Efficiency of the communication tools - Quality of the technical support	Wang, 2003; Sun et al., 2008; Wu et al., 2010.

Level of Interaction	- Collaboration activities during the course - Interactions among peers - Interactions between peers and mentors - Diversity in assessment - Instructor response timeliness	Wang, 2003; Frederiksen et al., 2006; Wu et al, 2010; Spears, 2012.
E-learning pace	- Opportunity of self-regulation regarding time, place, and learning processes - Comparison between the estimated time to complete the activities with the effective effort and time required - Self-paced learning opportunities	Sun et al., 2008; Paechter et al., 2010.
Overall experience	- Comparison between the experience and learning expectations - Opportunity to suggest the same experience to other colleagues - Opportunity to repeat the experience with the focus on different contents - The possibility to have a wider access to the contents considering a payment fee	Bollinger and Martindale, 2004; So and Brush, 2008.

Table 2. Synthetic view on the Big Data analysis (techniques, data, and results)

		LEARNING ANALYTICS TECHNIQUES		
		Sentiment Analysis	**Classification**	**Clustering**
LMS DATA SOURCE	**Discussion forums**	Calculate the sentiment of each post	Label each post with the dominant LS dimension	Extract hot topics for each set of labelled (classified) forum posts
	Questionnaire	Calculate the sentiment of each answer	Not applicable *(the sections of the questionnaire correspond exactly to the classification topics)*	Extract hot topics for each LS dimension

Table 3. The questionnaire for the LS, and the associated topics for the classification

Section	Key LS Dimensions	Questions in the final questionnaire for course evaluation	Topic
1	Course structure and contents	Provide your opinion about the structure of the course and the architecture of each online module. You can refer to the clarity of contents and learning objectives, coherence between them, typologies of assessment, etc.	Content
2	Technology and Support	Describe your experience with the system and technical support. You can refer to the usability of the platform, the easiness to access to the services, the efficiency of the communication tools, the quality of the technical support, etc.	Technology
3	Level of Interaction	Describe the quality of interaction within the community (peers, mentors, tutors, managers). You can refer to the collaboration activities happened during the course, the interactions among peers, between peers and mentors, etc.	Interaction
4	E-learning pace	Describe how you consider the pace at which the course advances, compatibly with your motivation level and learning goal. You can refer to the comparison between the estimated time to complete the course activities and assignments with the effective workload and time really required, etc.	Pace
5	Overall experience (method, technologies, teachers, and contents)	Provide an overall judgment on the learning experience done. You can refer to the opportunity to suggest the same experience to other colleagues, or to repeat the experience with the focus on different contents, or to the possibility to have a wider access to the contents considering a payment fee, etc.	Experience

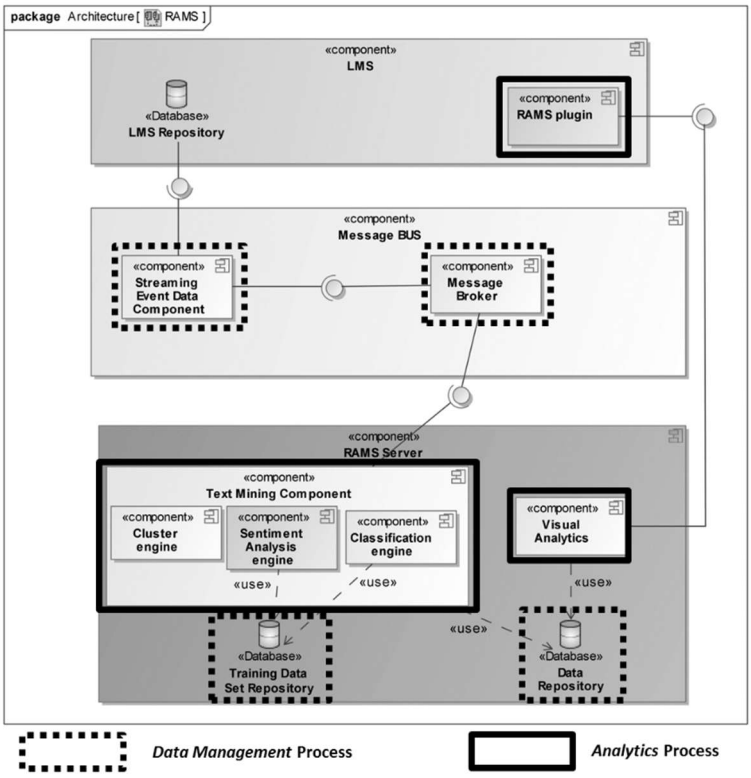

Figure 1. RAMS Architecture

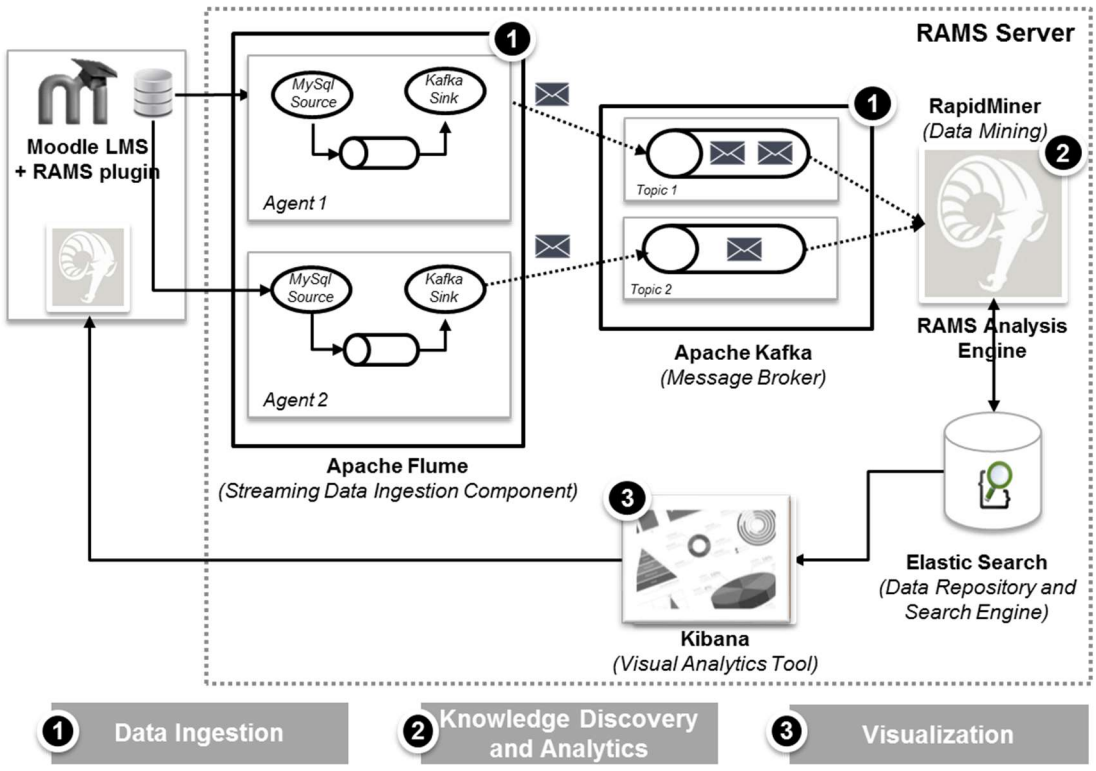

Figure 2. Big Data infrastructure (Open source software components and data flows)

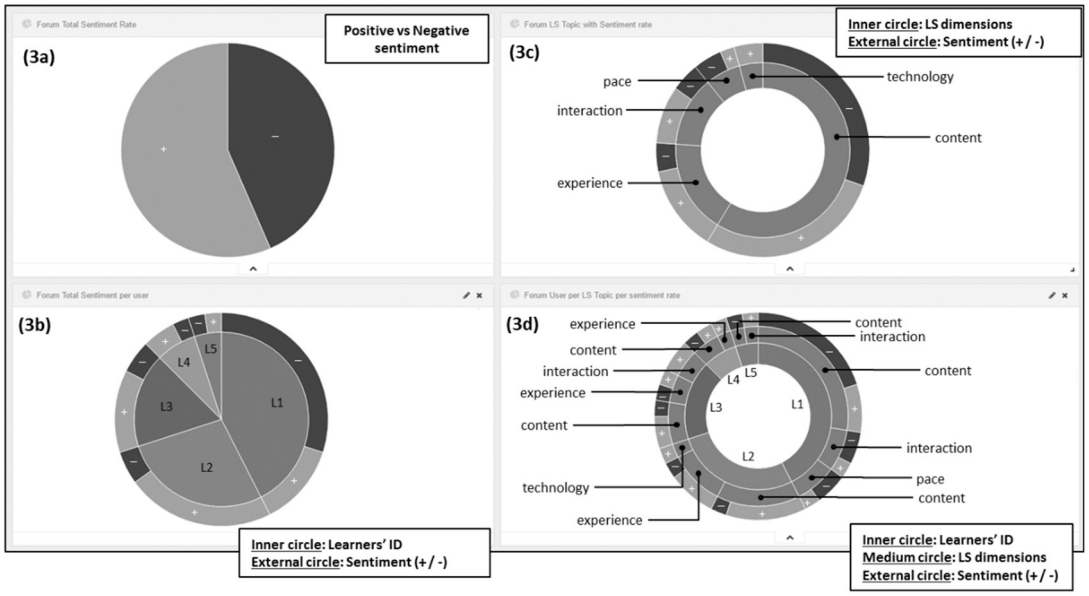

Figure 3. The forum sentiment dashboard

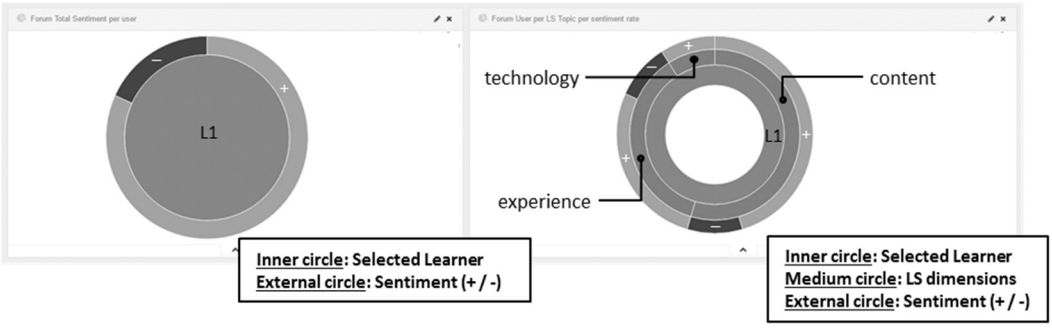

Figure 4. Drill-down results (learner-specific) on the forum sentiment dashboard

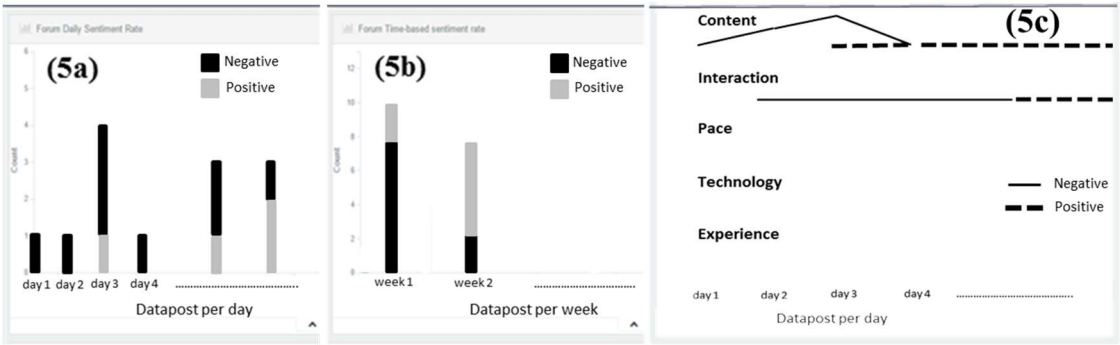

Figure 5. Time-based sentiment rate of the forum dashboard

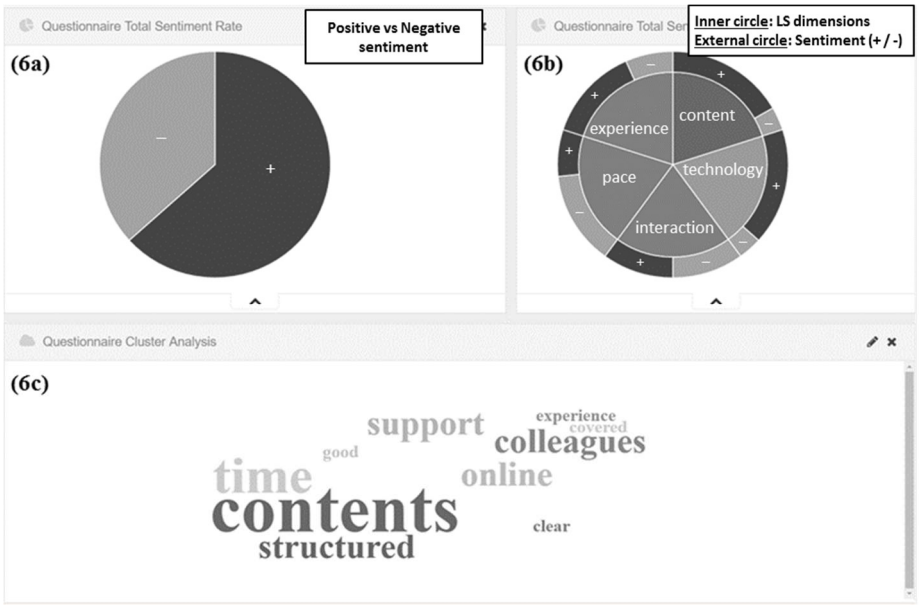

Figure 6. Clusters and sentiment rate of the questionnaire dashboard

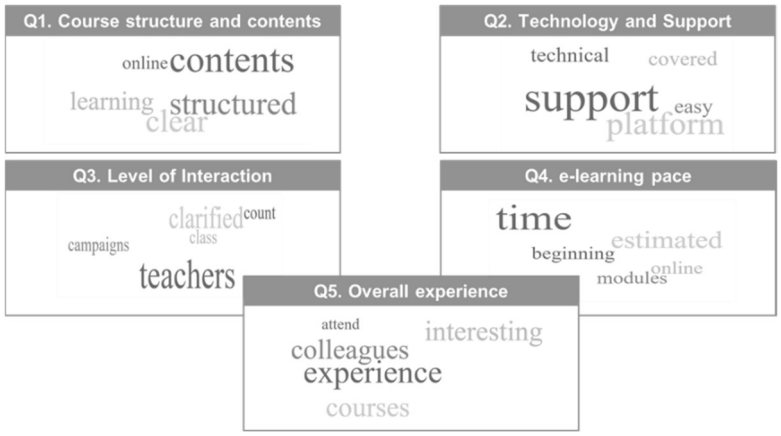

Figure 7. Clustering analysis applied to each questionnaire section

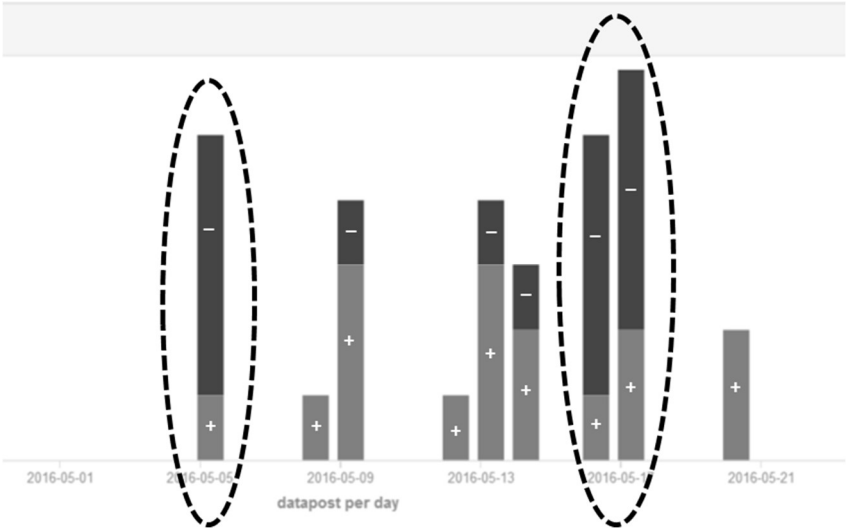

Figure 8. Daily sentiment rate and “critical situations” highlighted by RAMS