Journal of Universal Computer Science, vol. 25, no. 5 (2019), 541-568
submitted: 1/10/18, accepted: 20/5/19, appeared: 28/5/19

A Learning Ecosystem for Linemen Training based on Big Data Components and Learning Analytics

Guillermo Santamaría-Bonfil
(CONACYT-INEEL, Instituto Nacional de Electricidad y
Energías Limpias, Cuernavaca, México
guillermo.santamaria@ineel.mx)

Guillermo Escobedo-Briones
(INEEL, Instituto Nacional de Electricidad y
Energías Limpias, Cuernavaca, México
gescobedo@ineel.mx)

Miguel Pérez-Ramírez
(INEEL, Instituto Nacional de Electricidad y
Energías Limpias, Cuernavaca, México
mperez@ineel.mx)

Gustavo Arroyo-Figueroa
(INEEL, Instituto Nacional de Electricidad y
Energías Limpias, Cuernavaca, México
garroyo@ineel.mx)

Abstract: Linemen training is mandatory, complex, and hazardous. Electronic technologies, such as virtual reality or learning management systems, have been used to improve such training, however these lack of interoperability, scalability, and do not exploit trace data generated by users in these systems. In this paper we present our ongoing work on developing a Learning Ecosystem for Training Linemen in Maintenance Maneuvers using the Experience API standard, Big Data components, and Learning Analytics. The paper describes the architecture of the ecosystem, elaborates on collecting learning experiences and emotional states, and applies analytics for the exploitation of both, legacy and new data. In the former, we exploit legacy e-Learning data for building a Domain model using Text Mining and unsupervised clustering algorithms. In the latter we explore self-reports capabilities for gathering educational support content, and assessing students emotional states. Results show that, a suitable domain model for personalizing maneuvers linemen training path can be built from legacy text data straightforwardly. Regarding self reports, promising results were obtained for tracking emotional states and collecting educational support material, nevertheless, more work around linemen training is required.

Key Words: Big Data, Experience API, Learning Analytics, Learning Ecosystems, Text Mining.

Category: L.1.0, L.2.5, L.3.0, L.3.5

1 Introduction

Electricity utilities are companies in charge of electricity generation, transmission and distribution as well as power infrastructure maintenance. In the distribution process, maintenance is carried out by highly skilled workers called *Linemen* in accordance to specific maintenance maneuvers. Due to safety, technical, and business reasons, the effective training of this staff is mandatory [Vanrobayes and Roussel, 2017]. However, the Linemen Maintenance Training (LMT) is complex since it involves physical preparation, knowledge on physics and electricity, theoretical and physical knowledge of maintenance maneuvers execution, as well as hygiene and safety industrial rules [U.S. Department of Energy, 2017, Caha et al., 2017]. All of these skills are put together during the execution of maintenance maneuvers to successfully realize it while avoiding injuries and risk.

Motivated by the need of having well-trained operators, the electrical industry has ventured in the usage of e-learning technologies since several decades ago. Among the technologies that have been adopted stand the non-immersive Virtual Reality Training System (VRTS), Augmented and Mixed Reality, and Learning Management System (LMS) [Ayala-García et al., 2016]. However, data generated by these systems is stored in its own database using non-standardized ad hoc formats. In consequence, the exploitation of data in favor of personalized education using Machine Learning (ML) has been hampered by the lack of interoperability among systems and heterogeneous storage systems.

On the other hand, non-traditional learning data sources have recently emerged. It is composed by a variety of sources such as e-learning platforms, websites, social networks, forums, mobile applications, self reports, and so on. Such systems represent a new vein of research opportunities for the exploitation of massive amounts of data for building adaptive learning systems. For instance, it is well known that social networks generate massive amounts of data in several formats (e.g. images, videos, text, emojis, etc). Thus, an appropriate technological framework is required to collect and analyze such data for the sake of personalized education.

Consequently we propose a Learning Ecosystem (LE) for LMT, a combination of a multitude of technologies and support resources which exploits data to provide individualized learning within the ecosystem [Hruska et al., 2015]. This framework is based on the usage of an educational standard called experience API (xAPI) [Kevan and Ryan, 2016b], and Big Data envisioned components. The former will allow formal and informal distributed learning experiences to be standardized for its collection [Kevan and Ryan, 2016b], whereas the latter provides the tools to ingest, process, and manage both legacy databases and large volumes of information corresponding to learning experiences. Second, we propose to employ Learning Analytics (LA) to exploit educational data and learners' traces collections to regulate and enhance education [Peña-Ayala, 2018].

More precisely, LA will be used for building components for providing personalized assistance to the LMT. In particular, in this work we emphasize the usage of Text Mining (TM) for building components such as the Domain Model, and the analysis of informal learning experiences to provide support services for tailoring personalized LMT. It is worth mentioning that, while building the domain model from legacy e-Learning is straightforward, SR processing and analysis requires big data components. Either way, a LE shall exploit both, legacy and new e-learning data.

In the rest of this paper we argue that by integrating an xAPI-based ecosystem using a Big Data architecture along with LA, is possible to build a LE for LMT from heterogeneous sources. Thus, first we present the related work regarding e-Learning in LMT and Big Data. Latter the functional architecture for the proposed LE is presented. Then, experiments on building the domain model and informal learning experience analysis are presented and discussed. Finally, conclusions and the future research for developing such system is discussed.

2 Related Work

In accordance to Blooms taxonomy [Hodaie et al., 2018], the learning objectives required by LMT are the cognitive, affective, and psychomotor domains. The first requires a learner to obtain specific mental skills to solve intellectual problems, whereas the second demands considering emotions and their effect in trainees performance. Its worth mentioning that, the psychomotor domain in LMT has been kept within real-world camp training due to the large associated costs in hardware and software. Thus, in the remaining of this section we briefly present the related work for technologies already adopted in the LMT for addressing cognitive and affective learning objectives. These are VRTS, LMS, and Intelligent Tutoring Systems (ITS). Next, the related work for Big Data used for education, Learning Analytics, the xAPI standard, and the usage of Self Reports (SR) are presented. Finally, the Learning Ecosystem concept and related works are discussed.

2.1 Legacy e-Learning in Power Grid Maintenance Training

In the last decade, computer-based tools have been used in LMT. From these, the most used have been VRTS, which can be Non-Immersive, Augmented or Mixed Reality [Ayala-García et al., 2016]. The former offers more realistic environments and interactions than the latter, however, the development costs and time are lower for Non-Immersive than Augmented and Mixed VRTS. These have shown to be successful in LMT by reducing linemen accidents, and, for economic reasons, the vast majority of VRTS for LMT are Non-Immersive [Ayala-García et al., 2016]. In short, a non-immersive VRTS consists in 3-D environments in which maneuvers are elaborated in a step-based fashion. Instructions are provided in written and spoken forms, then users interact with the virtual environment via computer mouse or keyboard. Unfortunately, VRTS have focused on the fidelity of real world tools and environments, and immediate pedagogic support, whereas trace data generated from these systems have only been recently exploited [Hernández et al., 2016a, Hernández et al., 2017, Santamaría-Bonfil et al., 2017] in favor of personalized tutoring.

Another technological venue in utilities staff training is the usage of LMS. LMS store and present educational content associated with courses. It also collect interactions data between users enrolled in courses and the educational materials. Further, by providing synchronous and asynchronous knowledge acquisition, it increases training availability and allows reaching a broader audience. Using web technologies and e-learning for LMT, provides readily available tools for senior employees to formally document critical processes into means that are already adopted by the upcoming generations [Reder, 2006]. Even some utilities have introduced an LMS as part of their personnel training roadmap [Islas et al., 2007]. Hence, e-learning environments for utilities personnel training using LMS along with Shareable Content Objects Repositories have been proposed [Reyes et al., 2012, Argotte et al., 2011a]. Nevertheless, data logs generated from users activities and its exploitation in favor of personalized education has been oversighted.

ITS are intended to support and improve the learning process within a selected knowledge area accordingly to individual learner needs. The classical ITS setup uses four modules: Knowledge Domain, Student, Tutor, and Interface. The first defines the components of a specific knowledge area; the second maps individuals knowledge, misconceptions, and behaviors, for providing specific instruction; the third is responsible for guiding the learning path by selecting the appropriate instructions and content, whereas the fourth presents, supports, and collects interaction between students and the ITS. ITS proposals for LMT have considered a blended-training strategy, affective estimation and animated pedagogical agents, and open learner models [Hernández et al., 2016b]. These proposals have focused on the cognitive and affective domains. Notwithstanding, none of these have considered the technological requirements to generate, integrate, and exploit data from training systems. Furthermore, once trainee finishes a maneuver, a new one is assigned following a LMT curricula whose follow up depends entirely on the instructors subjective judgment and expertise. For instance in [Argotte et al., 2011b, Argotte et al., 2011a], the domain model was represented using concept structure maps, thus a Tutor may determine the sequencing of content or courses. Nevertheless, such maps were entirely defined by domain experts which not only bias models but also requires large amounts of human work.

2.2 New e-Learning Venues

In accordance to [De Mauro et al., 2015], Big Data is any information asset characterized by a high volume, speed, and variety, which requires a specific technological framework and analytic methods for turning it into knowledge.

In the field of education, Big Data usage is growing at a fast pace. For instance, governments are beginning to generate reports of Big Data potentials in education, embracing these for the exploitation of educational data, and even envisioning how these technologies will be incorporated in higher educations institutions in the years to come [Eynon, 2013, Reyes, 2015, Birjali et al., 2018]. Nevertheless, there are few works which propose or discuss the technical aspects required by an Educational Big data framework. For instance, in [Michalik et al., 2014] the technical details for proposing a big data architecture suitable for universities are reviewed. The resulting architecture is composed by traditional business intelligence tools, Not Only Structured Query Language (NoSQL) databases, the Apache Hadoop^{TM} framework, Hadoop Distributed File System (HDFS) as the filesystem (i.e. how data is stored and retrieved), and Hive as the data warehouse. Similarly, in [Birjali et al., 2018] authors discuss an architecture for building educational systems using HDFS and Hive, along with Apache Flume for collecting and transferring distributed log files from educational systems. In [Santoso and Yulia, 2017] a big data warehouse using Hadoop and traditional databases is proposed for ingestion, staging, management, and data presentation platform for higher education institutes to provide support for decision making. However, the author fail in detailing which components of the Hadoop framework shall be employed.

The concept of Big Data is deeply intertwined with that of Learning Analytics. The latter is consider a multidisciplinary paradigm for wrangling, manipulating, modeling, and visualizing data from different educational sources to address: learners behaviors and performance, measuring social impact in learning, students' performance prediction, emotional states assessing, identifying student's learning strategies, provide decision making tools for educational stakeholders, and so on [Peña-Ayala, 2018, Kitto et al., 2015, Manso-Vazquez et al., 2018]. A key aspect of LA is that it goes beyond traditional statistics and analytic activities by including educational stakeholders in the overall process for making sense of information, coming to decisions and policies based on data.

Fig. 1 shows the general framework employed by LA [Santamaría-Bonfil, 2018], which closely resembles the data science process [O'Neil and Schutt, 2013]. It starts with a real-world problem (e.g. predicting students performance), data is then collected and manipulated to conform a data set suitable for ML modelling. In parallel, an Exploratory Data Analysis (EDA) is performed to visualize patterns and identify the most adequate ML tools to exploit these. Afterwards, findings are communicated and validated by educational stakeholders. This part is critical in LA since it allows stakeholders to ponder the usage of ML models in decision making. In accordance to the former, data-based products are built (e.g. model for predicting students performance) and launched, then, the process iterates into new products or into refinement of previous ones.

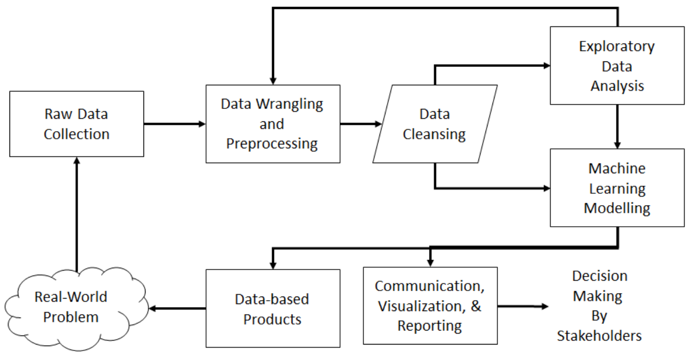


Figure 1: *Steps involved in the data science process used for performing Learning Analytics.*

2.2.1 Learning Ecosystems

Smart devices require big data technologies to communicate, consolidate, and store personal information to provide personalized feedback [Gros, 2016]. In this sense, the xAPI standard ensures that students' learning experiences data can be collected and shared between virtual learning environments and systems. This standardization is possible by transforming messages into *Activity Statements* (AS).

Altogether, Big Data infrastructure, the usage of the xAPI standard, and its exploitation for tailoring personalized instruction conforms a Learning Ecosystem [Hruska et al., 2015]. Some recent LE examples are a generic framework based on xAPI and GIFT [Hruska et al., 2015], a Live Fire Training LE proposed by the U.S. Army [Durlach et al., 2015], the Connected Learning Analytics toolkit for harnessing and exploitation of social media data [Kitto et al., 2015], Transmedia Learning [Raybourn, 2014], and a xAPI-based framework for collecting and monitoring Self Regulated Learning [Manso-Vazquez et al., 2018]. In particular, the last two rely on the usage of Self Reports (SR) for self monitoring, tutor monitoring, measure participants attitudes towards training, and so on. In this sense, the xAPI standard opens an opportunity not only for reporting learning activities, but also for for gathering and exploiting emotional self reports such as the Positive And Negative Affect Schedule [Egloff et al., 2003] or the Discrete Emotions Questionnaire (DEQ) [Harmon-Jones et al., 2016].

For reasons that will be later clear we delve DEQ more. It is a self reporting instrument (i.e. questionnaire) for measuring eight discrete different emotional states (i.e. anger, disgust, fear, anxiety, sadness, happiness, relaxation, and desire). This questionnaire requires any subject undergoing an emotional experience (e.g. watching a video) to rate words associated with the recognized emotions using a scale which ranges from 1 (i.e. not at all) to 7 (i.e. an extreme amount). Under one specific stimuli (e.g. anger due to failing in memorizing a sequence of steps) or a mixture of emotions (e.g. anger and anxiety for not been able to pass the course) the questionnaire will, presumably, capture the corresponding emotional reaction.

3 Proposed Learning Ecosystem

In the following we detail the LE for LMT shown in Fig. 2. We divide the LE in three layers: a) Systems and Data Sources, b) Ingest, Process and Storage, and c) ITS-like Components. The first layer includes, but is not limited to, several data sources such as LMS, Traditional Training, VRTS, SR, mobile applications, and social media. These record trace data from trainees and communicate it through a Message-Oriented Middleware (MOM). The second layer considers the Big Data components such as the Learning Record Store (LRS), Apache Spark, and HDFS. Any learning experience already mapped as an AS will be immediately stored in the LRS, whereas other message formats will be handled by the Apache Spark. The third layer considers LA data exploitation and the ITS-like components. Data products include, but are not limited to, the basic ITS modules, emotional assessment module, and other educational support components such as recommendation engines for course selection or content promotion [Peña-Ayala, 2018]. For the time being, we only focus on the Knowledge model and Educational Support.

In the following we detail the proposed Big Data architecture, the xAPI standard and AS, data sources particularly SR of informal learning and emotions, and how LA is used to enhance LMT personalized instruction.

3.1 Big Data Architecture

The communication among systems, storage devices, and intelligent learning components within the LE is vital. A component that allows retrieving and exchanging information between these is the MOM. In short, a MOM is a connectivity software designed for building large scale distributed systems providing synchronous and asynchronous messaging services to the systems while maintaining its independence [Curry, 2005]. In some cases such as LMS or SR, the

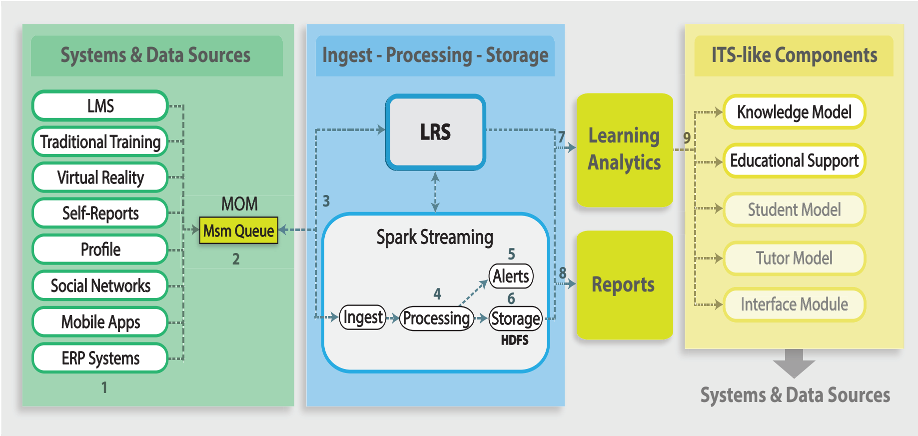


Figure 2: *Proposed Learning Ecosystem for Linemen training. From left to right, Systems and Data Sources, Data Ingestion and Storage components, and ITS-like components derived are shown, respectively.*

systems have already adopted the xAPI standard [Berking, 2016]. However, other systems such as VRTS, traditional training, or social networks may not comply with the xAPI standard. In both cases, a middleware is required. Thus, besides the communication part, the MOM will determine the proper processing/storage system in accordance to messages quantity and format.

In the case of AS messages, these will be delivered directly to an LRS. In accordance to its core specification, an LRS is a cloud-based system which is in charge of storing and retrieving learning data exclusively formatted as xAPI statements [Berking, 2016]. In its most basic setup only storing and retrieving functions are considered. However, depending on the LRS provider these may also include tools and dashboards to visualize, combine, aggregate, and manipulate AS data. Further, LRS already provide enough bandwidth capacity to ingest large amounts of AS statements.

In the case messages are not in the xAPI format, these would be processed and ingested by Apache Spark$^{\mathrm{TM}}$. Apache Spark is a computing system developed in the Scala language, focused on processing data in parallel across a cluster by working in-memory. Unlike Hadoop which requires intermediate data writing steps to HDFS, Spark processes data in RAM using a concept known as *Resilient Distributed Dataset* which avoids writing results each time a data is processed. In comparison with Hadoop, Spark is 100 times faster in operations such as the number of disk accesses per second and memory bandwidth utilization, whereas Hadoop consumes less memory [Samadi et al., 2017]. It also supports SQL queries, streaming data, and ML applications, among other features. Thus, Spark will be used in the LE as follows: if messages can be parsed immediately as an xAPI statement, these will be parsed and streamed to the LRS for its storage; if messages requires more processing or simply can be turned into an AS (e.g. videos, 3D models or environments), then these will be saved into the HDFS.

The HDFS is a scalable and portable system written in Java used as a distributed storage system which supports parallel access. Although similar to other existing distributed file systems, HDFS offers several advantages over the former. For instance, HDFS is ideal for storing large files, suitable for handling large data sets, and highly fault tolerant due of its low cost hardware design [Samadi et al., 2017]. In this sense, HDFS is meant to be used in the proposed LE for storing large data elements such as LMT videos or 3D components as scenarios or models; also as a redundant database.

3.1.1 xAPI Activity Statements

xAPI activity statements are syntactical similar to English syntax. In its simplest form, AS are composed by

$$\textit{an } \textbf{\textit{Actor}}, \textit{ a } \textbf{\textit{Verb}}, \textit{ and an } \textbf{\textit{Object}}. \tag{1}$$

The *Actor*, corresponds to a unique id associated to a specific subject (e.g. Lola the trainee). The *Verb*, such as in any language, classifies an actor's activity using a unique internationalized resource identifier. *Objects* can be of several types as long as it contains an *id* property whose value complies with the Uniform Resource Identifier (URI) scheme for unambiguous identification [Kevan and Ryan, 2016a, Manso-Vazquez et al., 2018]. For example, if *Lola the trainee* (Actor) *executes* (Verb) the *step 1 from maneuver 1 through the VRTS* (Object) will generate an AS as the one shown in Snippet 3.

It is worth mentioning that, AS statements are coded using the JavaScript Object Notation (JSON) for fast storage and retrieving. Also, AS can be completed using other fields provided by the xAPI standard such as *Context* (complementary information about the action), *Results* (grades and duration of the action), or *Extensions* (data which does not fit in other fields) [Manso-Vazquez et al., 2018]. In particular, the Extension field can be employed within the Object, Context, or Results fields as a property.

3.2 Data Sources & Systems

VRTS are already been used by utilities to carry out LMT and other staff training [Ayala-García et al., 2016]. In particular, we describe the VRTS tool called ALEn3D, a non-immersive VRTS used by the main Mexican utility for its LMT.

```
"actor": {
        "objectType": "Agent",
        "name": "Lola the trainee",
        "mbox": "mailto:lola.trainee@utility.gov"},
"verb": {
        "id": "http://adlnet.gov/expapi/verbs/executed",
        "display": {
        "en-US": "executed"}},
"object": {
        "objectType": "Activity",
        "id": "VRTS:Man1-Step1",
        "definition": {
                "name": {
                "en-US": "Step 1 in Maneuver 1"}}}
```

Figure 3: *The JSON pseudo code for an Activity Statement Example.*

In such tool, maintenance maneuvers are presented to the trainee using 3D power systems environments where these are safely elaborated step by step. Instructions about the goal, the steps and sub-steps, tools and materials required, and hygiene and safety rules involved, are provided written and in audio to the trainee. The trainee practices the maneuver in the VRTS until the instructor is satisfied by the trainee's observed proficiency. All instructions related to each maneuver presented through the VRTS are stored within a local database. For this particular case, the former is a relational database.

SR are not employed for LMT, hence, a proper self reporting tool is needed. For this endeavor a bookmarklet, a small software stored as a bookmark in a web browser, can be employed. Such SR bookmarklet can produce xAPI statements allowing to capture learning experiences (e.g. read, comment, watch, tweet, etc) from online content such as a video, a forum, an open course, and so on. Since it already produces xAPI statements, its data can be immediately reported to the LRS.

On the other hand, the SR bookmarklet can be employed for SR emotional states using the *Extension* xAPI field discussed earlier. This field allows defining new attributes using a combination of a key (or property) and a value using the URI scheme. The key value will be determined using emotions distinguished by the DEQ [Harmon-Jones et al., 2016] which are: anger, disgust, fear, anxiety, sadness, happiness, relaxation, and desire. Snippet 4 shows the extension field using DEQ. For simplicity, we employ the extension property within the object for SR emotional states. Thus, using the same example as above, if *Lola the trainee* (Actor) during the *execution* (Verb) of the *step 1 from maneuver 1 through the VRTS* (Object) felt *happiness*, the object xAPI component will be changed as follows:

```
        "object":{
                "objectType": "Activity",
        "id": "VRTS:Man1-Step1",
        "definition":{
                "name":{
                        "en-US": "Step 1 in Maneuver 1"
                        }}
        "extensions":{
                "http://id.tincanapi.com/extension/DEQ":
                "happiness"
                }}
```

Figure 4: *The JSON pseudo code of the AS Extension field required for emotional SR.*

3.3 Learning Analytics

LA is a multidisciplinary paradigm, among the disciplines employed by it stands TM, Natural Language Processing (NLP), Web Data Mining, EDA, ML, among others [Peña-Ayala, 2018]. We focus this work on the usage of TM, EDA with unsupervised algorithms for explore two venues: building a Domain model using texts of instructions in steps and sub-steps for each maintenance maneuvers, and explore SR usage using TM processing techniques and visualization tools.

3.3.1 Text Mining for LMT Domain Modelling

Instructions text used by VRTS already convey experts knowledge in maintenance maneuvers, thus, this can be used for building an LMT Domain model using Hierarchical Clustering (HC) algorithms [Rossi and Rezende, 2011, Fujihara and Batres, 2016]. Using maneuvers instructions texts, similarity groups can be built at several grain levels, from finer levels such as tools/materials (e.g. screw) and specific actions (e.g.tighten a screw), to coarser ones such as steps and courses. For the time being, we focus on the coarser domain model which corresponds to courses. However, before we can build it, data must be transformed into a suitable form for ML algorithms. First, a preprocessing step to standardize text is carried on. This consists in retrieving all instructions texts from the VRTS data base, and consolidate these into documents i.e. all steps which conform a maintenance maneuver is a document. Next, we extract documents' features by putting text to lowercase, removing special characters (e.g. Spanish accents), punctuation, and common words (i.e. stop words), and stemming, to reduce words to its root form to simplify words aggregation [Kwartler, 2017]. Once data is processed we use Bag of Words (BoW), a model which represents every word (or groups of words) as a unique document features [Kwartler, 2017]. The result of this is called a Document Term Matrix (DTM). In this, each row represents a document, whereas its columns represent features (i.e. stemmed words); each feature value corresponds to the appearance frequency of these in the document.

Using the DTM, a HC algorithm can be built. Typically, hierarchies are built using a bottom-up strategy, which implies that at the beginning each document is treated as a singleton cluster. As the structure grows up pairs of the most similar/dissimilar documents merge (or agglomerate) into coarser clusters until all of them have been merged into a single one. A HC requires two components: 1) a distance function to compare similarity between documents, and 2) linkage criteria to determine from where distance is computed. The former traditionally uses measures such as mathematical norms (e.g. 1-norm, 2-norm, p-norm) or the Pearson correlation. The latter determines how clusters will merge i.e. clusters merge by comparing their most similar members (single-linkage), by comparing the two most dissimilar documents (complete-linkage), the center of the clusters (mean or average-linkage), and so on [Manning et al., 2009]. In this work we employed 4 clustering algorithms: Hierarchical Agglomerative Clustering (HAC), the DIvisive ANAlysis Clustering (DIANA), K-Means, and Partition Around Medoids (PAM). The first two are hierarchical clustering, but the second is a top-down method whereas the first is bottom-up. The other two are partitioning methods which separate the feature space into non-overlapping clusters and are used as benchmark. Further, while both minimizes the within-class distance to the cluster centroid, K-means calculates the centroid as the middle point in the cluster whereas PAM selects the most central instance as the centroid.

Additionally, the resulting hierarchical structure depend on the number of clusters which is unknown. Further, is context-dependent and often several solutions are equally good from a theoretical point of view. Thus, to get a deeper insight on the stability of the resulting clusterings we applied three quality measures. These are the Connectivity Index ($Conn$), the Dunn Index ($Dunn$), and the Silhouette Width Index (Sil) [Handl et al., 2005]. The first, shows how connected are clusters as determined by the k-nearest neighbors, it takes values $Conn \geq 0$ and should be minimized. The second represents the smallest distance ratio between, observations not in the same cluster to the largest intra-cluster distance, it takes values $Dunn \geq 0$ and should be maximized. The third, averages each observations confidence degree for a particular clustering assignment. It measures ranges from $-1 \geq Sil \leq 1$, where a good clustering ≈ 1 and poor one ≈ -1.

3.3.2 Text Mining for Processing Self Reports

In order to exploit SR in favor of personalized learning, either from informal learning or emotional tracking, first we require to understand data gathered by the LRS. In the case of informal learning SR, we use TM to extract domains from websites to use them as ids. We focused on websites domains to constrain the analysis of web resources reported in informal learning SR. These along with the *Results* field of the AS specification, will reveal students preferences for web educational resources such as MOOC platforms, video lectures, blogs, forums, and so on. Since the *Results* field is readily available, we preprocessed the xAPI Object field by converting text to lowercase, and stemming it.

In the case of emotional SR, a simple TM preprocess is applied to the content of the Extension field within the Object field. This along with the *Results* field of the AS specification, can be used as a proxy of students emotional state as well as its evolution. Likewise informal learning SR, we normalized text from the Extensions field by converting it to lowercase, remove noise (i.e. map poorly-written emotions to its correct form), and stemming it.

3.3.3 EDA for Self Reports & Domain Modelling

In accordance to [Pearson, 2018], EDA is an approach to analyze data sets aided by visualization tools for discovering important patterns or structures within the data sets. These visualization tools may be *exploratory* to unveil and anatomize the content of a data set or *explanatory* for conveying findings to others [Pearson, 2018]. In this sense, the purpose of EDA for SR will be exploratory whereas for Domain modelling is explanatory.

Regarding SR, we decided to employ basic visualizations such as bar charts and boxplots. The reasons for using these are two-fold: first, data is generally better explored at a coarse granularity level; second, bar charts are effective in summarizing the relative frequencies, magnitude differences, and displaying integer-valued numerical data [Pearson, 2018] (such as ratings of informal learning SR), while boxplots are useful in displaying summary statistics (i.e. mean and standard deviation), study data distribution, and supplement multivariate displays. Hence, in the case of informal SR, bar charts allow to visualize which websites are the most visited, whereas boxplots allows to assess how good or bad were informal learning experiences regarding their ratings. Likewise, for emotional SR, bar charts will allow to get insight about how emotions evolved as the course progressed, while boxplots will allow instructors to assess how intense were the reported emotions. For instance, in the case of informal learning, using visualizations an instructor may determine websites value and even discuss it with colleagues for further exploitation, or, in the case of emotional SR, use it as a "thermometer" of how compelled are students with the class in accordance to his/her experience.

Regarding the domain modelling, the resulting clusters of a HC algorithm are visualized using a hierarchical tree structure called Dendogram. Such structure is useful to discern, not only similar objects, but also their clusters relationships. In this case, a dendogram will display in the x-axis documents and clusters, whereas on the y-axis the height (i.e. dissimilarity) between clusters. Thus, the dendogram will show how maintenance maneuvers, as characterized by their texts (i.e. actions and materials), shall be grouped not only locally (i.e. most similar maneuvers) but also globally (i.e. relationship between maneuvers clusters). For instance, using this knowledge domain structure an instructor (human or artificial) may personalize maneuvers learning path (i.e. which maneuvers a student shall learn) for each student.

4 Experimentation

Here we present the experimentation performed 1) for the exploitation of legacy e-Learning for building a Domain model, and 2) for pattern discovery in data from informal learning experiences and emotional SR. We detail first the experimental configuration and software used for the LE. Then, results for building the domain model for courses, and SR exploratory analysis are presented. Finally, limitations of the approach are discussed. It is important to notice that, data and processes from companies in the electric power industry, either operational or personnel training, is highly sensible, confidential, and access restricted. Hence, while we were able to collect data from linemen legacy e-Learning systems, SR have to show its value, before and not after, it can be brought to LMT training sessions. Thus, for the SR exploratory analysis a proof-of-concept is used.

4.1 Legacy e-Learning Experimental Setup

We process and extract information from a legacy VRTS software used by the main Mexican utility for LMT in medium-tension maneuvers [Hernández et al., 2017]. The medium-tension LMT program covers 43 maintenance maneuvers which range from rescuing an injured linemen due to an electric shock to the replacement of several pole structures, in every case with energized lines. All LA experimentation was carried out using R language along with the RStudio™.

The VRTS stores all maneuvers information within an informix® database, which is located at the LE data sources layer (recall Fig. 2). Since this data does not come from any learning experience source, the MOM will connect it directly with the LA component. However, if data from legacy e-Learning is too large or cannot be stored in a relational database (i.e. videos about a maneuver's execution), the MOM will redirect data, first, to Spark streaming component for its ingestion, processing, and proper storage. Once in the LA component, using standard SQL queries we retrieved the information corresponding to the written instruction texts for each step and sub step of each maintenance maneuver. Next, data is processed by the TM techniques earlier presented to obtain its DTM form. The shortest maneuver texts contains 296 words, whereas the longest maneuver is composed by 1600 words, both after stemming.

4.2 SR experimental setup

As already mentioned, due to utilities restrictions before we are allowed to gather data from LMT we most prove its value. Thus, a proof-of-concept is carried out. The latter consists in the usage of the proposed LE platform by students during a theoretical course, such as the one received by linemen during their theoretical formation (e.g. electrical concepts or maneuvers training). The LE must be useful for collecting educational material which will be employed as educational support content and tracking trainees emotional states for instructors assessment. Therefore, we employed as a proxy for this analysis, data from 9 postgraduate students (master and PhD) enrolled in a Machine Learning course (part of a master degree program [INEEL, 2018]) which was carried out from January 2018 to June 2018. Although postgraduate students backgrounds are rather different to those of linemen, it has been stated more than a decade ago, that new linemen generations have already adopted new e-Learning technologies and are currently learning from a computer [Reder, 2006].

For gathering and performing SR, Rustici Software LLC bookmarklet was employed [Initiative, 2018]. This is a plug-and-play bookmarklet which allows to capture 4 basic learning experiences (i.e. experience, read, bookmark, and tweet) and is located at the LE data source layer. It also allows to rate AS using a scale which ranges from 1 (lowest) to 5 (highest), and personalized them using tags. These tags, allows to include additional information such as Context information or declare emotional states. On the other hand, the component which consolidates all AS is the LRS located at the LE ingest, process and storage layer. Once data is stored at the LRS, it can be retrieved by the LA component for furthering analysis and modelling. From a pool of LRS providers, the Watershed^{TM} LRS was selected in accordance to the cost-benefit (unlimited AS reports and users for free). The SR bookmarklet, and AS statements stored by the LRS are shown on the left and right part of Fig. 5, respectively. The latter corresponds to SR time series; interestingly while the overall SR time series (above) depicts two large peaks at the end of the course, individual interactions (below) reveals that such peaks are due to an specific individual. However, these visualizations neither distinguish between informal learning and emotional SR, nor analyzes ratings or any additional information.

Using the bookmarklet, SR were generated following a set of instructions, one for informal learning experiences and another for emotional states. In both cases, at the beginning of the course, students were presented with SR concepts and technology. Particularly, for the DEQ case, students were provided with Table 6 of DEQ seminal paper [Harmon-Jones et al., 2016] for them to identify each emotional state. Also, we carried out several customizations to the instrument as recommended by DEQ authors [Harmon-Jones et al., 2016]: 1) reduced DEQ words items to the broad emotions categories, and 2) constrain scale between 1 $\leq$

$Rate \leq 5$. Such customizations were established for avoid increasing students' cognitive load, and allowing several emotions to be reported at once [Harmon-Jones et al., 2016]. Afterwards, students were notified that a minimum quota of reports (one report for each course hour) will be required, and that SR would constitute 20% (i.e. 10% each) of their final course grade.

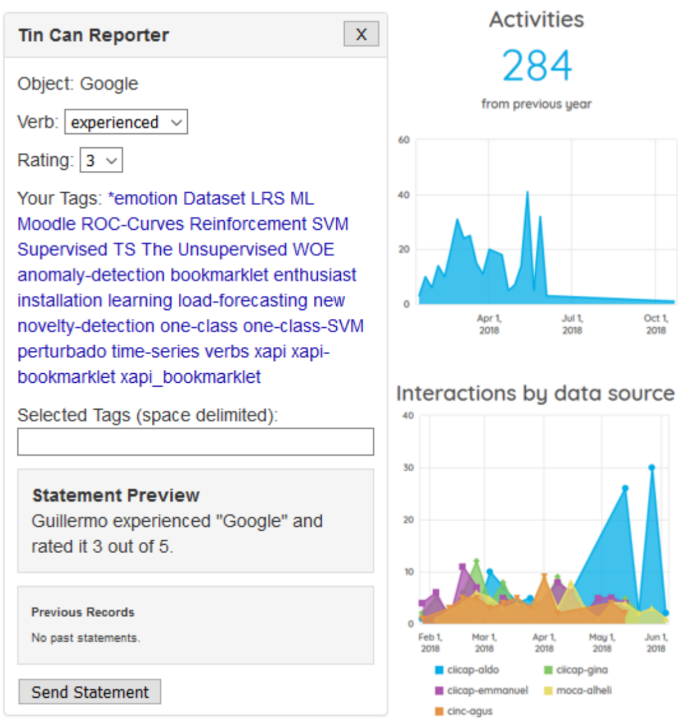


Figure 5: *Self-Reporting using the xAPI Bookmarklet. On the left, the xAPI bookmarklet used to report is shown, on the right, screen shots from LRS visualizations of the overall (above) and individuals (below) SR.*

4.2.1 Instructions for Self Reporting Informal Learning

Informal SR will be elaborated, out from the classroom, during any learning session related to the course. While undergoing this learning session, the student have to identify any material she/he had used for learning (e.g. video, blog, forum, etc.) that is *worthy* to report. In this sense, a material is considered worthy if students appraises it as important, useless, clarifying, confusing, or so on, regarding a course topic. This subjective appraising shall employ the following scale: 1) not at all helpful, 2) slightly helpful, 3) somewhat helpful, 4) very helpful, and 5)extremely helpful. Thus, for instance, a fast paced and confusing material should be rated using a low value (i.e. $rate \leq 3$) whereas a document which explains a topic very neatly should have a high rate (i.e. $rate > 3$).

4.2.2 Instructions for Self Reporting Emotional States

Emotional states SR will be elaborated, out from the classroom, during any learning session related to the course. At any moment of this learning session, the student had to report his/her emotional state employing one or more of the emotions defined in Table 6 of DEQ seminal paper. Such emotional SR had to be accompanied by its corresponding rating (i.e. how intense was/were) emotions. This subjective emotional appraising shall employ the following scale: 1) not at all intense, 2) slightly intense, 3) somewhat intense, 4) very intense, and 5)extremely intense.

4.3 Results

Here we present and discuss results for the aforementioned experiments setup.

4.3.1 LMT Domain Hierarchy

After maintenance maneuvers texts are posed as a DTM, we applied the aforementioned clustering algorithms. To analyze the stability of the resulting clusters for each algorithm, we applied *Conn*, *Dunn*, *and Sil* quality measures for $K = 2, \ldots, 10$ groups whose results are shown in Fig. 6: HAC is shown in black, K-means in red, PAM in green, and DIANA in blue.

Observe that, in accordance to *Sil* and *Conn*, the best number of clusters is 2, whereas the *Dunn* index is maximized for 3. Also, hierarchical clustering methods mostly outperformed their partitioning counterparts for these quality measures. Particularly, K-means obtained the overall worst results, whereas HAC slightly outperformed DIANA. Although for *Sil* and *Conn* a stability monotone decrease is depicted, for *Dunn* index, both hierarchical algorithms as well as PAM achieved a very high *Dunn* value for 8 clusters. This is important, since a hierarchy with only two levels is rather useless for the purpose of personalizing instruction. Therefore, we decided to build a hierarchy using HAC with 8 clusters.

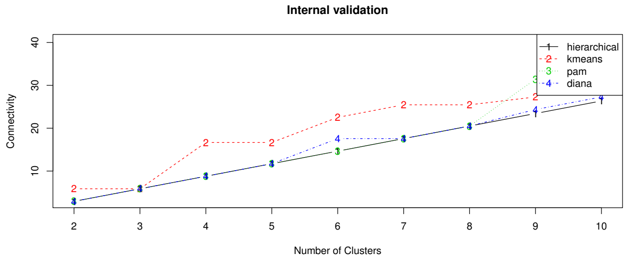
The resulting LMT Domain Hierarchy is presented in Fig. 7 showing the clusters within the red boxes. This structure reveals how maintenance maneuvers are related with each other, thus, a human or virtual tutor that is (located at the ITS-like Layer of the LE) can consume this hierarchy to select the best learning path for any lineman trainee. For instance, since hierarchy's height determine dissimilarity between maneuvers, an instructor could determine that smaller branches shall be taught first, i.e. maneuvers 30 and 33, followed by 26 and 42, and so on. Another learning path solution could be to taught first those clusters whose maneuvers are more similar with each other. Hence, the fifth cluster (from left to right) would be selected, been maneuvers 7 and 43 taught first, followed by 10, 12, and so on. Also, it is notorious the case of maneuver 1. Unlike the rest, this is considered a special linemen maneuver devoted to help and rescue an injured linemen. Thus, that this maneuver was separated from the rest in the resulting hierarchy, was not only expected but also desirable.

4.3.2 Exploring Self Reports

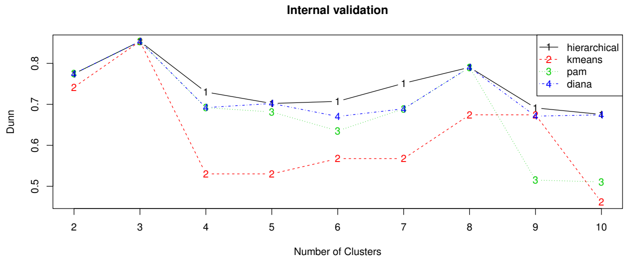
Analyzing informal learning and DEQ SR is thrilling. Both SR analyzes was carried out using time intervals of two weeks. Regarding informal learning SR, in Fig. 8 we present the objects (i.e. websites domain) reported by the students of the proof-of-concept course. However, since a myriad of website domains were reported by students, we decided to kept only the top 12 most reported. On Fig. 8(a) reports evolution across course duration are shown; the x-axis presents intervals dates whereas the y-axis the frequency of reports. On the other hand, Fig. 8(b) presents boxplots corresponding to websites rating for the top 12 website domains; boxplots are sorted from the most (top) to least (bottom) reported websites. Recall that, ratings convey student's subjective appraisal about the website resource with respect to a topic from the course. Observe that, YouTube is not only the most reported website but also it was visited throughout most of the course. Its content rating median is 4, although its rates ranged from slightly (i.e. 2) to extremely (i.e. 5) helpful. From all top sites it is the one with most dispersion. Following YouTube, it is the statistics department portal from Charles III University of Madrid (i.e. halweb.uc3m.es), the datacamp MOOC platform (which is famous for hosting several ML courses), and moodlecloud.com (a simple LMS employed for content management of the proof-of-concept course). These are clearly related to the topic of the course. Furthermore, notice that reports are prevalently neutral and positive.

The idea of collecting informal learning SR from students is to conform a repository of educational support content. First, employing time stamps altogether with tags of SR reports, we can match them with their correspondent course activity. Then, when a student is learning or solving a particular topic, an instructor (human or synthetic) may suggest a related list of web resources (sorted in accordance to their rates), in the same fashion as a recommendation engine. These recommendations shall take into consideration similar (in category [1]) resources as those previously visited by the student. Thus, for instance, if a

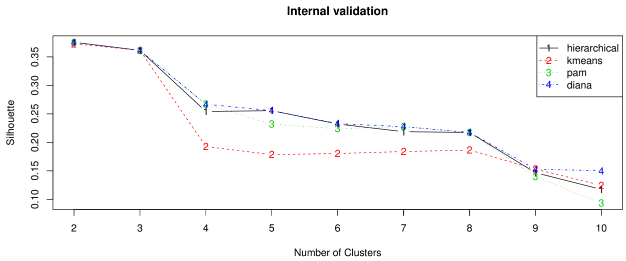
[1] `https://en.wikipedia.org/wiki/Category:Lists_of_websites`.



(a)



(b)



(c)

Figure 6: *Quality Measures results for 4 clustering algorithms. On (a) connectivity, on (b) the Dunn Index, and on (c) the average Silhouette Width.*

student prevalently carries on informal learning through video hosting websites and MOOC platforms, seeks information in forum websites (e.g. if websites have been somehow categorized (e.g. videos site, MOOC site, forum site, etc.) such suggestion

As for emotional SR, Fig. 9(a) presents the overall emotion reporting evolution across course duration, whereas Fig. 9(b) presents, within boxplots, emo-

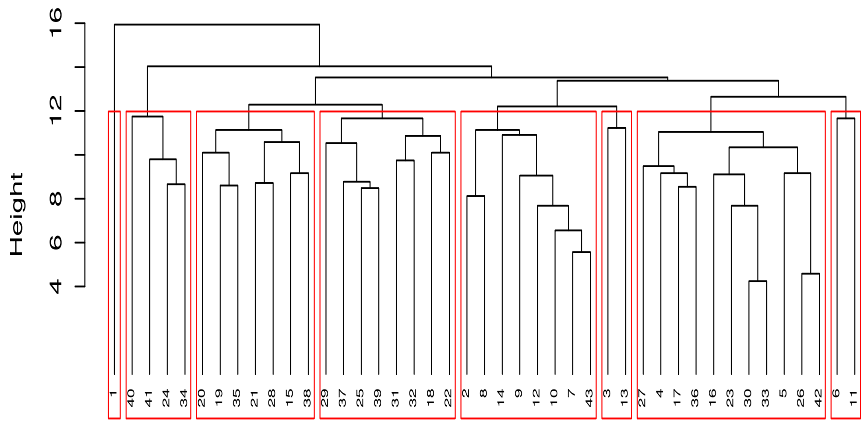
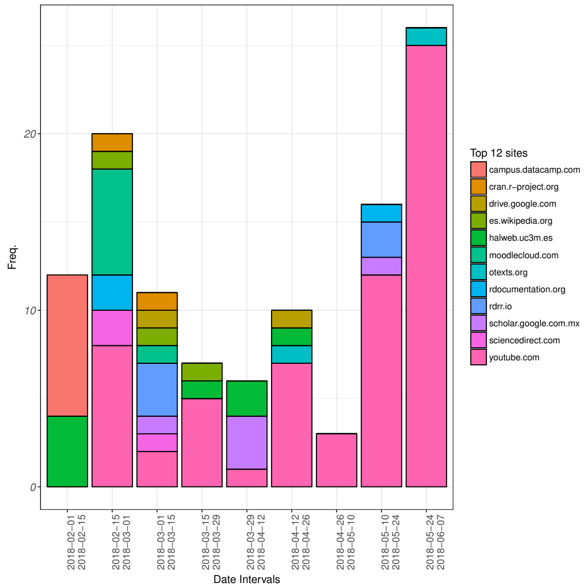
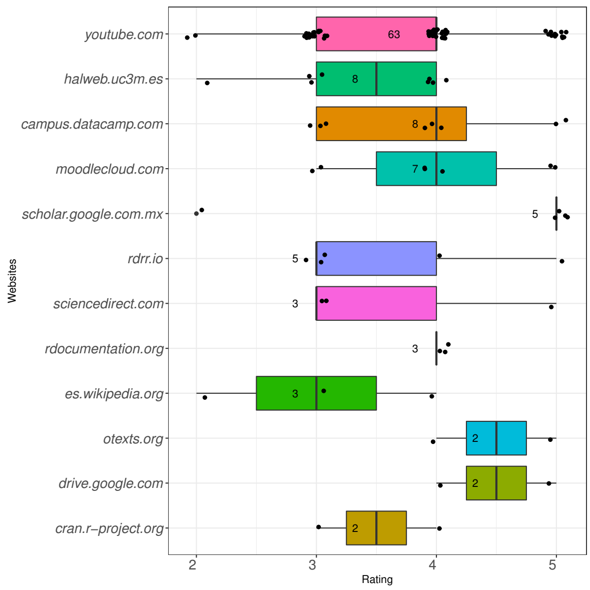


Figure 7: *Dendogram of the reconstructed LMT Domain Model using the HAC algorithm. The k selected clusters are displayed within red boxes.*

tions rating distributions during informal learning experiences. In the former, the x-axis presents the DEQ emotions and the y-axis the reported frequency; from top to bottom, from left to right, boxes display time as goes by. In the latter, boxplots are sorted from the most (top) to least (bottom) reported emotion; ratings shown at the x-axis, correspond to the subjective emotion intensity felt by student's during an informal learning experience. Notice that, in both figures the DEQ scale fear is absent. Regarding emotions self reporting across time, we can see that reports were more frequent at the early stages of the course, decreasing heavily at the end of it. Also, observe that some intervals were dominated by one specific emotion such as desire (March 1-15 and April 12-26), relaxation (Feb 15-March 1), or anxiety (Feb 1-15 and May 10-24), which were also the most popular emotions, respectively. Boxplots, on the other hand, shows popular, medium popular (happiness and anger), and unpopular (disgust, sadness) emotions distributions. The median for popular emotions was very intense while for the rest it ranged from somewhat to extreme intensities; overall, all emotions variances were contained within slightly to extreme intensities. Both, boxplots and time evolution bar charts, shows that some DEQ emotion scales are, in appearance, more related to informal learning experiences than others. For instance, the fear scale was not reported a single time. This coincides with DEQ authors claims about necessary customizations required for its usage and extension (e.g. modify DEQ scales). Notwithstanding, the here proposed DEQ customizations show that the DEQ instrument altogether with an xAPI Bookmarklet can be used to track and collect students emotional states without increasing significantly the cognitive load [Harmon-Jones et al., 2016]. Furthermore, given that emotional



(a)



(b)

Figure 8: *Informal Learning Experiences xAPI Statements. On (a) reporting evolution across time in two weeks intervals is shown, whereas (b) presents boxplots for the top 12 website domains ratings.*

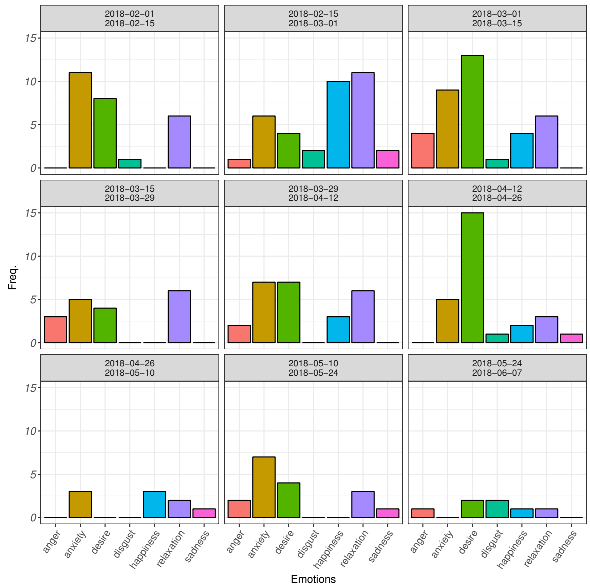
SR happens along with a learning activity, we can associate it with Objects (e.g. websites). In this sense, an instructor may provide any student with content which is not only related to specific topics but also with highly rated positive emotions (and even eliciting them).

4.4 Limitations

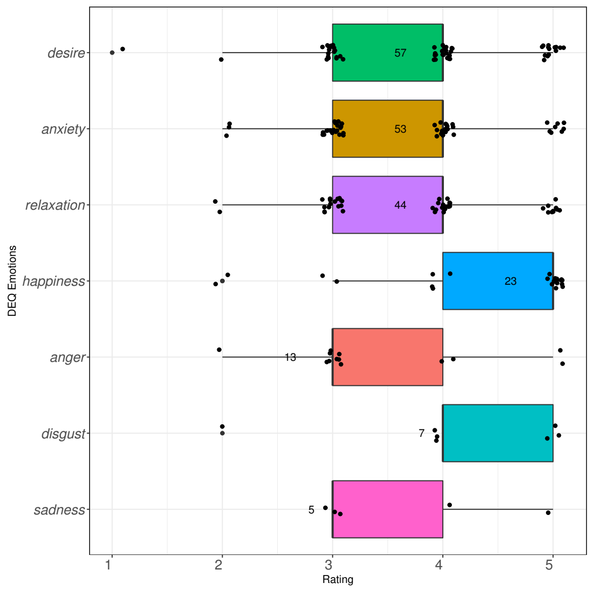
Although results are promising there are several limitations of the proposed framework. On one hand, for the Big Data architecture since both, LRS and the Spark framework support Machine Learning algorithms, for the time being it is unclear where will be deployed the Learning Analytics data-base products. In the same fashion, the exploitation of the constructed domain model, the support content repository, and tracking and assessment of emotions, requires an instructor (human or artificial intelligence-based) which will decide when and how to apply these data-based products. Further, it is required to deepen in the xAPI requirements to properly standardize the DEQ Extension, and to improve the bookmarklet for ease reporting (e.g. using check and combo boxes for selecting emotions and its intensities).

With respect to the experiments, there is large room of improvement. Nowadays SR is not been used in the LMT, thus linemen SR data does not exists. Hence, even while results obtained for the proof-of-concept are useful, it is necessary to adequate the proposed experiments along with utilities educational stakeholders for assessing its true value within the LMT domain. In the same venue, while DEQ fear scale was not reported for the proof-of-concept course, we should expect it to appear in the LMT context since most mistakes in this domain have deadly consequences. In the same fashion, since emotions were provided by students through the bookmarklet in written form rather than chosen from a fixed list, this lead to the appearance of new emotions (e.g. melancholy) or other somatic perceptions (e.g. being tired). For the time being, since non-standardized reported emotions represented barely 0.03% these were mapped to DEQ emotions. Also, DEQ was devised as an instrument for measuring discrete emotions while undergoing 1) an isolated emotional experience 1) in any given circumstance. In this sense, neither is designed to be employed consecutively (i.e. scoring every informal learning experience using the 32 words provided by the original DEQ questionnaire surely must be exhausting!) nor it is specific for learning experiences. Therefore, additional customization and emotional states (which are more specific to learning), are required.

Regarding technical aspects, the fact that the quality measures used to evaluate domain knowledge modelling showed that the best number of clusters were 2, reflects that the characterization of maneuvers by its text can be improved (BoW models are unable to capture word order and semantic meaning). Thus



(a)



(b)

Figure 9: *DEQ xAPI Statements. On (a) reporting evolution across time in two weeks intervals is shown, whereas (b) presents boxplots for emotions ratings. Notice that the fear scale was not reported.*

more advance TM models such as syntactic parsing, or neural network embeddings shall provide better results.

5 Conclusions & Future Work

In this work we presented a proposal for building a LE for LMT. This goes towards generating a standardized system for exploiting both, legacy and new e-Learning data sources, which will provide personalized instruction in the LMT domain. To date we have focused upon the software components and standards which are necessary for building the LE framework, as well as some LA applications namely Domain modelling, and the collection and exploration of informal learning and DEQ emotional SR. Results for the domain modelling are useful in determining a LMT learning path, however, these are considered to be improved using algorithms which can captured semantic meaning in maneuvers text descriptions. On the other hand findings about SR usage in the proof-of-concept course are encouraging. The next steps involve designing better experimental tools (i.e. improving the bookmarklet interface), standardize DEQ emotions within the xAPI specifications, and enlarging the population size by establishing contact with utilities managers responsible of carrying out the LMT courses. Finally, we will also integrate legacy and new VRTS into the LE, and through the collection of trainees trace data we expect to generate a student model which is able to assess trainees knowledge about LMT.

Acknowledgements

Author GSB thanks the Cátedra CONACYT program for supporting this research.

References

[Argotte et al., 2011a] Argotte, L., Arroyo-Figueroa, G., and Noguez, J. (2011a). SI-APRENDE: An Intelligent Learning System Based on SCORM Learning Objects for Training Power Systems Operators. In *Dev. Concepts Appl. Intell.*, volume 363, pages 33–38. Springer-Verlag Berlin Heidelberg.

[Argotte et al., 2011b] Argotte, L., Hernández, Y., and Arroyo-Figueroa, G. (2011b). Intelligent elearning system for training power systems operators. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6882 LNAI, pages 94–103.

[Ayala-García et al., 2016] Ayala-García, A., Galván-Bobadilla, I., Arroyo, G., Pérez-Ramírez, M., and Muñoz-Román, J. (2016). Virtual reality training system for maintenance and operation of high-voltage overhead power lines. *Virtual Reality*, 20(1):27–40.

[Berking, 2016] Berking, P. (2016). Choosing a learning record store (LRS). Technical report, Advanced Distributed Learning Initiative. Accessed: 2018-09-29.

[Birjali et al., 2018] Birjali, M., Beni-Hssane, A., and Erritali, M. (2018). Learning with Big Data Technology: The Future of Education. In *Proceedings of the Third International Afro-European Conference for Industrial Advancement AECIA 2016. AECIA 2016. Advances in Intelligent Systems and Computing*, volume 565, pages 209–217.

[Caha et al., 2017] Caha, V., Miletic, Z., Peric, S., and Raljevic, D. (2017). Live work training centres cooperation and development. In *2017 12Th International Conference on Live Maintenance (ICOLIM)*, pages 1–3.

[Curry, 2005] Curry, E. (2005). Message-Oriented Middleware. In Mahmoud, Q., editor, *1*, chapter 1, pages 1–28. John Wiley & Sons, Ltd, Chichester, UK, 1 edition.

[De Mauro et al., 2015] De Mauro, A., Greco, M., and Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *AIP Conference Proceedings*, volume 1644, pages 97–104.

[Durlach et al., 2015] Durlach, P., Washburn, N., and Regan, D. (2015). Putting Live Firing Range Data to Work Using the xAPI. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2015*, number 15019, pages 1–11.

[Egloff et al., 2003] Egloff, B., Schmukle, S., Burns, L., Kohlmann, C., and Hock, M. (2003). Facets of Dynamic Positive Affect: Differentiating Joy, Interest, and Activation in the Positive and Negative Affect Schedule (PANAS). *Journal of Personality and Social Psychology.*

[Eynon, 2013] Eynon, R. (2013). The rise of Big Data: What does it mean for education, technology, and media research? *Learning, Media and Technology*, 38(3):237–240.

[Fujihara and Batres, 2016] Fujihara, S. and Batres, R. (2016). A Micro-Genetic Algorithm for Ontology Class-Hierarchy Construction. *Int. J. Comput. Linguist. Appl.*, 7(1):51–65.

[Gros, 2016] Gros, B. (2016). The design of smart educational environments. *Smart Learning Environments*, 3(1):15.

[Handl et al., 2005] Handl, J., Knowles, J., and Kell, D. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

[Harmon-Jones et al., 2016] Harmon-Jones, C., Bastian, B., and Harmon-Jones, E. (2016). The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions. *PLOS ONE*, 11(8):e0159915.

[Hernández et al., 2016a] Hernández, Y., Cervantes-salgado, M., and Pérez, M. (2016a). Data-driven Construction of a Student Model using Bayesian networks in an Electrical Domain. In *MICAI 2016: Mexican International Conference on Artificial Intelligence*, pages 481–490.

[Hernández et al., 2016b] Hernández, Y., Pérez-Ramírez, M., Zatarain-Cabada, R., Barrón-Estrada, L., and Alor-Hernández, G. (2016b). Designing empathetic animated agents for a B-learning training environment within the electrical domain. *Educational Technology and Society*, 19(2):116–131.

[Hernández et al., 2017] Hernández, Y., Santamaría-Bonfil, G., and Pecero, V. (2017). Text Mining for Domain Structure Analysis in a Training System for Electrical Procedures. In *10th Workshop on Intelligent Learning Environments, WILE 2017*, pages 1–7.

[Hodaie et al., 2018] Hodaie, Z., Haladjian, J., and Bruegge, B. (2018). TUMA: Towards an Intelligent Tutoring System for Manual-Procedural Activities. In *Intelligent Tutoring Systems-14th International Conference*, volume 10858, pages 326–331. Springer International Publishing.

[Hruska et al., 2015] Hruska, M., Medford, A., and Murphy, J. (2015). Learning ecosystems using the generalized intelligent framework for tutoring (GIFT) and the experience API (xAPI). In *Proceedings of the 17th International Conference on Artificial Intelligent in Education*, volume 1432, pages 9–16.

[INEEL, 2018] INEEL (2018). Master In Energy Science Program. `https://www2.ineel.mx/posgrado/maestrias/mce.html`. Accessed: 2018-09-29.

[Initiative, 2018] Initiative, A. (2018). xAPI Bookmarklet. `https://xapi.com/bookmarklet/`. Accessed: 2018-09-29.

[Islas et al., 2007] Islas, E., Pérez, M., Rodriguez, G., Paredes, I., Ávila, I., and Mendoza, M. (2007). E-learning tools evaluation and roadmap development for an electrical utility. *Journal of Theoretical and Applied Electronic Commerce Research*, 2(1):63–75.

[Kevan and Ryan, 2016a] Kevan, J. and Ryan, P. (2016a). Experience API: Flexible, Decentralized and Activity-Centric Data Collection. *Technol. Knowl. Learn.*, 21(1):143–149.

[Kevan and Ryan, 2016b] Kevan, J. M. and Ryan, P. R. (2016b). Experience API: Flexible, Decentralized and Activity-Centric Data Collection. *Technol. Knowl. Learn.*, 21(1):143–149.

[Kitto et al., 2015] Kitto, K., Cross, S., Waters, Z., and Lupton, M. (2015). Learning analytics beyond the LMS: the Connected Learning Analytics Toolkit. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*, pages 11–15.

[Kwartler, 2017] Kwartler, T. (2017). *Text Mining in Practice with R.* John Wiley & Sons, Ltd, Chichester, UK.

[Manning et al., 2009] Manning, C., Raghavan, P., and Schutze, H. (2009). Hierarchical Clustering. In *Introd. to Inf. Retr.*, chapter 17, pages 321–345. Cambridge University Press, Cambridge.

[Manso-Vazquez et al., 2018] Manso-Vazquez, M., Caeiro-Rodriguez, M., and Llamas-Nistal, M. (2018). An xAPI Application Profile to Monitor Self-Regulated Learning Strategies. *IEEE Access*, 6:42467–42481.

[Michalik et al., 2014] Michalik, P., Štofa, J., and Zolotová, I. (2014). Concept definition for Big Data architecture in the education system. *SAMI 2014 - IEEE 12th Int. Symp. Appl. Mach. Intell. Informatics, Proc.*, pages 331–334.

[O'Neil and Schutt, 2013] O'Neil, C. and Schutt, R. (2013). *Doing Data Science: Straight Talk from the Frontline*, volume 1. O'Reilly.

[Pearson, 2018] Pearson, R. (2018). *Exploratory Data Analysis using R.* CRC PRess, 1 edition.

[Peña-Ayala, 2018] Peña-Ayala, A. (2018). Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8(3):1–29.

[Raybourn, 2014] Raybourn, E. (2014). A new paradigm for serious games: Transmedia learning for more effective training and education. *Journal of Computational Science*, 5(3):471–481.

[Reder, 2006] Reder, W. (2006). The technical talent challenge [workforce development]. *IEEE Power and Energy Magazine*, 4(1):32–39.

[Reyes et al., 2012] Reyes, A., Buen, P., Islas, E., Mart, R., and Jim, F. (2012). Intelligent Tutoring and Training Tools for the Electric Power Sector Developed at IIE. *Research in Computer Science*, 47:81–93.

[Reyes, 2015] Reyes, J. (2015). The skinny on big data in education: Learning analytics simplified. *TechTrends*, 59(2):75–80.

[Rossi and Rezende, 2011] Rossi, R. and Rezende, S. (2011). Building a topic hierarchy using the bag-of-related-words representation. *Proc. 11th ACM Symp. Doc. Eng. - DocEng '11*, page 195.

[Samadi et al., 2017] Samadi, Y., Zbakh, M., and Tadonki, C. (2017). Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks. *Concurr. Comput.*, (October 2017):1–13.

[Santamaría-Bonfil, 2018] Santamaría-Bonfil, G. (2018). Towards a learning ecosystem for linemen training. In *WILE 2018: 11th Workshop on Intelligent Learning Environments at MICAI 2018*, pages 1–8, ITESM Campus Guadalajara, Guadalajara, México.

[Santamaría-Bonfil et al., 2017] Santamaría-Bonfil, G., Hernández, Y., Pérez-Ramírez, M., and Arroyo-Figueroa, G. (2017). Bag of Errors : Automatic inference of a student model in an Electrical Training System. In *MICAI 2017: Mexican International Conference on Artificial Intelligence*, pages 1–15.

[Santoso and Yulia, 2017] Santoso, L. and Yulia (2017). Data Warehouse with Big Data Technology for Higher Education. *Procedia Comput. Sci.*, 124:93–99.

[U.S. Department of Energy, 2017] U.S. Department of Energy (2017). Transforming the Nation'S Electricity System the Second. Technical Report January, U.S. Department of Energy.

[Vanrobayes and Roussel, 2017] Vanrobayes, B. and Roussel, F. (2017). Live work training and skills development at RTE. In *2017 12th International Conference on Live Maintenance (ICOLIM)*, pages 1–3. IEEE.

Acronyms Lists

Power Grid

LMT Linemen Maintenance Training.

e-Learning & Education

AS Activity Statements.

DEQ Discrete Emotions Questionnaire.

ITS Intelligent Tutoring Systems.

LE Learning Ecosystem.

LMS Learning Management System.

SR Self Reports.

VRTS Virtual Reality Training System.

xAPI experience API.

Learning Analytics

Conn Connectivity Index.

Dunn Dunn Index.

Sil Silhouette Width Index.

BoW Bag of Words.

DIANA DIvisive ANAlysis Clustering.

DTM Document Term Matrix.

EDA Exploratory Data Analysis.

HAC Hierarchical Agglomerative Clustering.

HC Hierarchical Clustering.

LA Learning Analytics.

ML Machine Learning.

NLP Natural Language Processing.

PAM Partition Around Medoids.

TM Text Mining.

Big Data & Programming

HDFS Hadoop Distributed File System.

JSON JavaScript Object Notation.

LRS Learning Record Store.

MOM Message-Oriented Middleware.

NoSQL Not Only Structured Query Language.

URI Uniform Resource Identifier.