R&I

research.uoc.edu

APPLICATION OF A TEXT ANALYTICS MODEL FOR THE AUTOMATIC DISCOVERY AND CLASSIFICATION OF RESEARCH TRENDS IN E-LEARNING (THE UOC CASE)

José López Ruiz and Guillem Garcia Brustenga

eLearn Center, Universitat Oberta de Catalunya

Report justification

As a research and innovation centre specializing in e-learning, the UOC's eLearn Center (eLC) continuously analyses global sources of knowledge to detect and reflect on trends and events that transform online higher education and lifelong learning around the world. As a result of this horizon scanning, regular reports on trends in the field of e-learning are published, such as last year's *E-learning Research Report 2017* (*eLRR 17*), which requires its experts to carry out repetitive extraction, analysis, synthesis, codification and classification tasks on hundreds, and in some cases thousands, of bibliographic references. Given the systemic nature of this analysis process and the large volume of articles and other documentary sources needing regular reviewing, data analysis technologies and text analysis tools were thought to have the potential to adequately cover part of this laborious process.

Under this hypothesis and based on the aforementioned report on trends in current e-learning research produced by the centre, the eLC carried out this innovative project for the automatic extraction and classification of scientific articles. The developed model permits relevant information to be extracted quickly and easily from scientific articles on e-learning and automatically classified by the chosen parameters (main topic, subject category, research methodology and educational stage).

The project was carried out with the participation of the company MeaningCloud, a specialist in cloud text analytics, to design and apply an automatic classification model based on semantic rules of deep categorization.

The work was divided into two sections:

1. Running a pilot test to validate the applicability of the text analysis tools to the automatic classification of scientific articles.

2. Carrying out a test of the automatic classification using e-learning publications from 2018.

Introduction to the model

Analysis of scientific literature to detect trends in a discipline or specific area of research is a common practice for researchers wishing to identify initiatives driving change in their field. Such knowledge is also essential for making decisions in designing, reformulating or abandoning projects that respond to the interests of institutions. Reviewing scientific articles and other documentary sources to extract significant information is a long and laborious process, even when the information is structured, and it can represent an immense effort for the team of analysts, depending on the degree of detail proposed for the search. Therefore, some of the work stemming from knowledge management, such as information extraction, analysis and classification, might be a good candidate for automatic processing.

In this context, data analytics and text mining technology could play a key role in effectively covering some of the phases in the research process. Text mining techniques may be defined as the application of computing methods and techniques to textual data to identify relevant and intrinsic information and previously unknown knowledge (Do Prado and Ferneda, 2007).

The project pilot test was implemented from this perspective. It consisted of trial automatic extraction and categorization of information relevant to world e-learning research in hundreds of high-impact scientific articles. The report *E-Learning Research Report 2017*. *Analysis of the main topics in research indexed articles,* drawn up by the centre's researchers, was chosen to evaluate the effectiveness of the analytical solution, as the same publications were used in its production. The study identifies trends in which e-learning researchers focused their interest and directed research efforts and resources during the period of analysis. The codification and classification previously defined in the research study was used to prepare the text analysis in order to identify and subsequently classify key concepts, so that the automatic results could be compared to and evaluated against those obtained by the report's authors.

Preparation and execution of the pilot test

Methodology

Design of the model began in 2018, based on the following methodology:

- The test was based on 855 articles on e-learning extracted from the main scientific publication databases, Web of Science (WoS) and Scopus, and analysed in the *eLRR 17* study. As the report explains, this sample was obtained after screening for articles repeated in both databases or which failed to meet the literature search criteria.

- Approximately half of the total articles were chosen for the pilot test. These were publications that contained the most frequent categories and topics, as this favoured the model's training and subsequent evaluation. The training sample consisted of around 63% of the chosen articles.

- Using the training sample, three rule-based deep categorization models were developed, which covered the taxonomies proposed for each of the levels of text analysis:

    - Stage of education
    - Methodology used
    - Thematic classification

- These models, following the habitual training, validation and test methodology, were matched against the test sample, which consisted of the rest of the articles (37%). This allowed the precision and real coverage of the three models to be calculated. These data were also matched against all the cases selected for the pilot test, so as to compare the similarity of the automatic classification with the manual classification carried out by the research group that produced the 2017 report.

- One of the questions raised while developing the classification models was whether the body of the article should be used for the classification or just the title and abstract. To answer this, a dual validation was carried out, comparing article title and abstract with title, abstract and complete text. It was concluded that analysis of only the title and abstract was more precise than the complete text (less than 70% relevance), hence it was decided to use only this part of the publications for the automatic classification of categories.

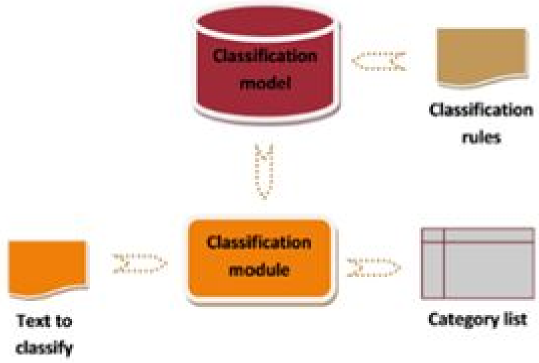
Classification rules

Text mining was used for the automatic extraction and classification of relevant information from the articles. The tool used was the MeaningCloud deep categorization engine (Fig. 1), which incorporates advanced rule technology covering lexical, syntactic and semantic levels in natural language. Thus, the classification rules were based on:

- Morphological information: differentiating the form of lemma and marking terms with their grammatical category.
- Syntactic information.
- Semantic information through a specific, customizable ontology.

In addition, advanced expressions (regular, proximity and logical operators) were also incorporated, so that rules were based on the meaning of the expressions rather than literal forms.

Figure 1. Diagram of the text classification process



The pilot test training corpus was built using the following parameters:

- Stage of education: 7 categories - 60 classification rules.
- Methodology used: 6 categories - 135 classification rules.
- Thematic classification: 33 categories - 230 classification rules.

Results of the pilot test

The evaluation continued through different iterations matching the training sample previously tagged by the authors of the report against the defined classification models. The classification models were further developed in the light of these iterations. At the same time, these classification models were matched against the test sample. The aim was to calculate their true efficiency, and, using all the sample articles, to extract conclusions from the comparison of the automatic and manually tagged classifications.
Comparison was carried out for each of the categories (thematic, methodology and stage of education) for title+abstract (see classification tables 1-3) and the whole body of the article. The precision and recall (*or exhaustiveness*) of the results, understood as a measure of the "usefulness" of the values returned in the search (*precision*) and how complete they were (*recall*) were measured in each comparison.

The tables comparing the frequencies in the classification models for stage of education, research methodology and article topics (see *eLRR 17*) are given below. The analysis was conducted on title and abstract for all pilot test articles (training and test). The comparison was based on applying the automatic classification model vs no model (UOC tagging).

Table 1. Results of the classification with the stage of education category: comparison of frequencies

Categories	Freq. without model	Freq. with model	Difference
K12Secondary	*8.82%*	*9.31%*	*0.49%*
Undergraduate	*58.58%*	*56.13%*	*–2.45%*
Vocational	*0.98%*	*0.98%*	*0.00%*
Postgraduate	*4.90%*	*3.43%*	*–1.47%*
Professional	*12.50%*	*13.24%*	*0.74%*
Other	*14.22%*	*16.91%*	*2.70%*

Table 2. Results of the classification with the methodology category: comparison of frequencies

Categories	Freq. without model	Freq. with model	Difference
Design	*10.25%*	*11.00%*	*0.75%*
Evaluation	*16.00%*	*15.50%*	*–0.50%*
Descriptive	*41.50%*	*39.00%*	*–2.50%*
Review	*9.75%*	*9.75%*	*0.00%*
CaseStudy	*8.25%*	*9.50%*	*1.25%*
Experimental	*14.25%*	*12.00%*	*–2.25%*
Other	*0.00%*	*3.25%*	*3.25%*

Table 3. Results of the classification with the thematic category: comparison of frequencies

Categories	Freq. without model	Freq. with model	Difference
Adoption	*11.00%*	*12.28%*	*1.28%*
Assessment	*7.42%*	*8.44%*	*1.02%*
Design Evaluation	*27.11%*	*28.90%*	*1.79%*
Disciplinary Research	*8.70%*	*10.23%*	*1.53%*
Innovation	*18.93%*	*17.14%*	*–1.79%*
Instructional Design	*8.44%*	*6.91%*	*–1.53%*
Interaction	*2.56%*	*3.32%*	*0.77%*
Learning Analytics	*7.67%*	*6.65%*	*–1.02%*
Student Performance	*3.07%*	*2.81%*	*–0.26%*
StudentPsychoVariables	*2.81%*	*1.28%*	*–1.53%*
TeacherRole	*2.30%*	*2.05%*	*–0.26%*

Based on the results of the categories examined, it was concluded that the **analysis of article title and abstract was sufficient to determine precisely each classification, without having to use the complete text**, hence terminological analysis of these parts was recommended and the latter was rejected.

Thus, given the taxonomies established in the report, using these shorter sections of publication would appear to be more useful in searching for key concepts. In the comparison of global percentages, the frequencies obtained for each category when extracted manually were very similar to those extracted using the automatic model (standard deviation less than 2%), initially validating the applicability these types of analytical solutions to knowledge management.

The analysis of the test articles also included experimentation with another data mining technique (clustering), used to divide data into groups of similar objects. The method was used to discover new relevant topics (organized into subtopics) similar to the article's main topic (see Table 4). This was an option because the model used here permits levels of relevance to be associated with the categories obtained in the analysis and previously unknown relations between them to be discovered.

Table 4. Results of the multi-category test

No. articles analysed	Minimum relevance threshold for title + abstract	Minimum relevance threshold for complete article
89	*30%*	*10%*
	Mean no. of categories obtained from title + abstract	Mean no. of categories obtained from complete article
	2	5

In 93.07% of cases, the category assigned by the engine in the text subtopic or multicategory classification was considered relevant. Considering these results, **complete article analysis becomes more recommendable for detecting and extracting relevant e-learning subtopics.**

Automatic classification of 2018 publications

Once the viability of these types of data analysis tools for the detection and classification of significant concurring terms in the articles had been demonstrated in the pilot test, it was decided to complete the model with all the taxonomies in the *eLRR 17* report and refine the tested ones until they were all covered. The aim was to fully conclude it in order to execute a complete automatic classification with the information from research articles on e-learning from 2018. Thus, the following descriptors were differentiated for each article:

- Main topic and its thematic category.
- Educational stages in which the research is set.
- Research methodology used in each publication.

The whole process was carried out using the same previously developed framework of thematic codifications and classifications as a reference, while excluding the work of reviewing, updating and identifying new topics not categorized in the previous report. Thus, unlike *eLRR 17*, which was a research study, the data in the following tables and graphs should not in any way be considered exact for defining trends, although they help to establish an idea of the state of research in e-learning in 2018.

The data below, together with the overall project results, help us confirm the benefits of using text mining to quickly process large volumes of information with which to draw precise conclusions to support the arguments in our studies on e-learning.

Results

Sample

A total of 1,043 articles on e-learning published in WoS and Scopus in 2018 were processed in the automatic analysis. The articles were obtained using the same literature search criteria as in *eLRR 17*.

Tables

Table 5. Classification of the top 20 topics from 2018 by volume of frequencies

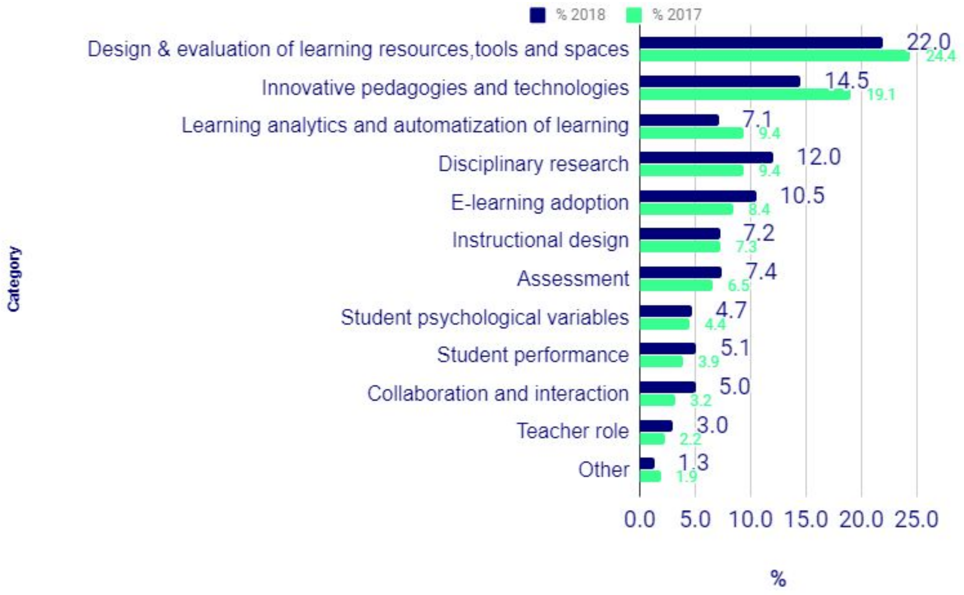
Position in 2018	Thematic category	Description	Variation since 2017	Frequency in 2018	% total category in 2018
1	Assessment	Type of assessment facilitated by digital environments (rubrics, e-assessment, training assessment, badges).	↑1	66	6.3
2	Mobile learning	Mobile and ubiquitous learning-based models.	↓1	55	5.3
3	MOOC	MOOC design, implementation and assessment (effectiveness, satisfaction).	0	55	5.3
4	Platform	Design, implementation, assessment and selection of e-learning platforms (LMS, VLE).	↑3	44	4.2
5	E-learning adoption	Implementation of distance courses where none previously	0	42	4.0

		existed (comparison with face-to-face learning, assessment of perception and efficiency).			
6	E-health	Publications on e-learning theory and practice in the fields of health and health promotion among the general population.	↑12	38	3.6
7	Training	Vocational e-learning in different subjects (health, business and enterprise, human resources and lifelong learning).	↑10	37	3.5
8	Communication	Publications on necessary communication, in its broadest sense (networks, language styles, tools, etc.) for online teaching and learning.	↑25	29	2.8
9	Gamification	Application of games and gaming strategies to improve learning (gamification, serious games).	↓3	29	2.8
10	Virtual reality	Virtual reality-based methodologies (simulations, virtual patients, virtual worlds, 3D).	↓6	28	2.7
11	Blended learning	Publications that assess the design of transition experiences from face-to-face to blended education, combining face-to-face and online methodologies.	↑8	26	2.5
12	Adaptive learning	Customized teaching and learning initiatives, adapted learning paths and adaptive feedback.	0	25	2.4
13	Acceptance	Studies that measure acceptance in the adoption of e-learning initiatives and components (courses, tools, platforms, etc.) by the parties involved.	↑12	24	2.3
14	Teacher	Study of the teacher's role and action in e-learning.	↑14	23	2.2
15	Student performance	Impact factors on students' performance in the learning	↑16	23	2.2

		activity. Systems and mechanisms to measure and assess them.			
16	Instructional design	Instructional, didactic sequence, activity and course design.	↓7	22	2.1
17	Course assessment	Design, implementation and assessment of specific online courses.	↓9	22	2.1
18	E-learning success factors	Assessment of variables that measure success and effectiveness in the integration of e-learning systems.	↑22	21	2.0
19	Social media	Design and introduction of social media applications for social interaction-based learning.	↓4	21	2.0
20	Remote laboratory	Design, development and resources regarding virtual or remote experimentation spaces for e-learning.	↑7	21	2.0

Graphics

*Figure 2. Comparison of yearly frequency by thematic category**



**Excluding 46 references which the system could not associate with any of the thematic categories structured in* eLRR 17*.*

Figure 3. Comparison of yearly frequency by methodology

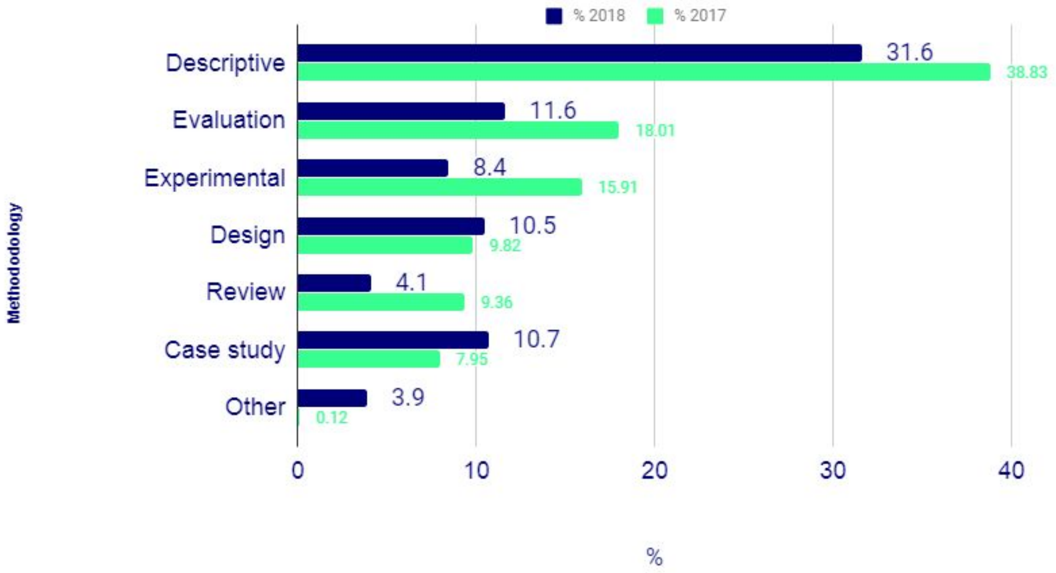
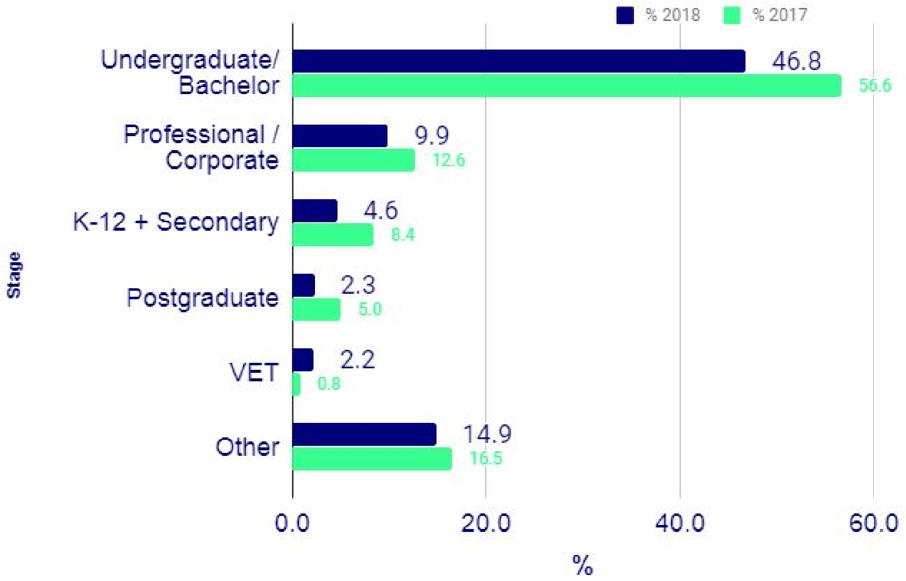


Figure 4. Comparison of yearly frequency by stage of education



Classification by subtopic

As mentioned in the pilot test section, because the deep categorization models provide a fragment-level analysis, they permit the main topic as well as subtopics to be extracted from a text, sorted by order of relevance. Thus, a "multicategory" classification is established by extracting the main topic in each article to then extract topics or subtopics to which the former may be significantly related. This characteristic could prove useful for discovering thematic correlations that help provide more in-depth knowledge of the context of a specific research study and, together with the application of other computational techniques, the configuration of the complete scientific domains.

By way of example, as shown in the following table, in three articles whose main topic was "Training", the predominant theme detected (with over 50% relevance) was "e-health". A detailed analysis of these three articles shows that they discuss research into professional training in the clinical field. The inclusion of subtopics might help indicate different relevant labels with which article content could be categorized.

Table 6. Sorting of multicategory topics

1st relevant or main topic	2nd relevant topic	No. of articles
Active learning	Assessment	**3**
Adaptive learning	Student performance	**3**
Assessment	Course design	**6**
	Course evaluation	**5**
	MOOC	**4**
	Communication	**4**
	Satisfaction	**3**
	Platform	**3**
	Mobile learning	**3**
	Social media	**3**
	Skills	**3**
Course evaluation	Assessment	**4**

	Blended learning	3
Course design	Student performance	3
E-learning success factors	Platform	3
E-health	Communication	3
E-learning adoption	Acceptance	3
	Course Design	3
Gamification	Assessment	4
	Skills	3
	Student performance	3
	Virtual reality	3
Learning analytics	Student performance	3
Mobile learning	Social media	3
Platform	Communication	3
	Assessment	3
	Resources	3
Training	Teacher	4
	E-health	3

Rules

After completing the model with all the categories, the rules used were:

- Stage of education: 7 categories - 93 classification rules.
- Methodology used: 6 categories - 154 classification rules.
- Thematic classification: 67 categories - 380 classification rules.

Conclusions

The data obtained from the project demonstrate the viability of using text mining techniques and automatic classification rule models for detecting and classifying trends in a specific field of research. This is especially so when the topics in the field of study being explored have already been structured and categorized in previous research studies, as in this case, thereby facilitating enormously the work of collecting and classifying existing information. The use of data mining for such research purposes or business intelligence is not a new phenomenon; the techniques have been applied to numerous consolidated knowledge fields and disciplines for some time.

However, it should be borne in mind that the field of e-learning is not exactly consolidated, as the thematic category still does not exist in world scientific publication systems (Tibaná-Herrera et al., 2018). As these authors explain, it is a fragmented domain, where, for instance, in bibliometric terms, articles on e-learning hover between technical and educational approaches, mainly in publication categories related to social sciences and computing. Thus, in the area of e-learning, problems may arise in the initial search process and data collection.

With regard to processing content, it is important to note that, although the designed taxonomy facilitates automatic classification of the articles, it did not consider "machine language". This adds certain ambiguities and overlaps in the processing that could affect the exhaustiveness and precision of the results. If the quality of the classifications is to be improved, it is essential to teach the system situations of context in which it knows how to identify whether the final object of a social network analysis (SNA) is assessing learning in leadership, whether reference to deep learning is made in terms of artificial intelligence (AI) or competencies, or that behind tests of certain forms of teaching communication lies the analysis of a pedagogical strategy.

While consensus grows regarding the uniqueness of the field of e-learning, it might be a suitable context to continue completing and adjusting an automatic model for collecting and categorizing trends, such as the one tested in the UOC. Implementations could be oriented to updating and adapting its dictionary to automatic processing, with further actions to strengthen the data analysis techniques and add a cognitive layer to provide greater precision in its classifications. This would also help to discover new and significant descriptors inside and outside the subject in all the scientific literature.

Acknowledgements

MeaningCloud™ (https://www.meaningcloud.com) was used for the text analytics in developing and validating this pilot test.

References

[1] eLearn Center (2018). *E-Learning Research Report 2017. Analysis of the main topics in research indexed articles*. Barcelona: eLearn Center (UOC). Retrieved July 17, 2019 from http://hdl.handle.net/10609/75705.

[2] Do Prado, H.A., Ferneda, E. (2007), *Emerging Technologies of Text Mining: Techniques and Applications*. Hershey, New York: Information Science Reference.

[3] Precision and recall (n.d.). In *Wikipedia.* Retrieved July 17, 2019, from https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=913652928

[4] Berkhin P. (2006), A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (Ed.) *Grouping Multidimensional Data*. Berlin/Heidelberg: Springer.

[5] Tibaná-Herrera, G., Fernández-Bajón, M.T., De Moya-Anegón, F. (2018). Categorization of E-learning as an emerging discipline in the world publication system: a bibliometric study in SCOPUS. *International Journal of Educational Technology in Higher Education 15*, 21.

[6] Sharma, D., Kumar B., Chand, S. (2018). Trend Analysis in Machine Learning Research Using Text Mining. Greater Noida (UP), India: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN),* pp. 136-141.