IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON APPLICATIONS OF BIG DATA IN SOCIAL SCIENCES

Received December 5, 2018, accepted January 9, 2019, date of publication January 15, 2019, date of current version February 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2892987

A Review of the Analytics Techniques for an Efficient Management of Online Forums: An Architecture Proposal

JESÚS PERAL[1], ANTONIO FERRÁNDEZ[1], HIGINIO MORA[2], DAVID GIL[2], AND ERICK KAUFFMANN[3]

[1]Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain
[2]Department of Computer Technology and Computation, University of Alicante, 03690 Alicante, Spain
[3]School of Industrial Engineering, University of Costa Rica, San José 11501-2060, Costa Rica

Corresponding author: JESÚS PERAL (jperal@dlsi.ua.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness through the Project SEQUOIA-UA under Grant TIN2015-63502-C3-3-R, the Project RESCATA under Grant TIN2015-65100-R, and the Project PROMETEO/2018/089, and in part by the Spanish Research Agency (AEI) and the European Regional Development Fund (FEDER) through the Project CloudDriver4Industry under Grant TIN2017-89266-R.

ABSTRACT E-learning is a response to the new educational needs of society and an important development in information and communication technologies because it represents the future of the teaching and learning processes. However, this trend presents many challenges, such as the processing of online forums which generate a huge number of messages with an unordered structure and a great variety of topics. These forums provide an excellent platform for learning and connecting students of a subject but the difficulty of following and searching the vast volume of information that they generate may be counterproductive. The main goal of this paper is to review the approaches and techniques related to online courses in order to present a set of learning analytics techniques and a general architecture that solve the main challenges found in the state of the art by managing them in a more efficient way: 1) efficient tracking and monitoring of forums generated; 2) design of effective search mechanisms for questions and answers in the forums; and 3) extraction of relevant key performance indicators with the objective of carrying out an efficient management of online forums. In our proposal, natural language processing, clustering, information retrieval, question answering, and data mining techniques will be used.

INDEX TERMS Online forums, natural language processing, analytics techniques, KPIs, data mining.

I. INTRODUCTION

The methods of learning have experimented a tremendous change, especially due to the novelties in the technology as well as the requirements requested by society. We can definitely highlight the advent of the online frameworks in their diverse ways, Open and Distance Learning (ODL) [1], as the key factor. The inherent features of these sites offer all types of educational tools to many students around the world. Hence, a very significant change in the educational costs can be established from the more concrete elements (such as educational buildings) to the technological infrastructures that provide knowledge.

As a result, a series of terms are emerging that aim to identify the evolution of this process:

- e-learning (online education in which teachers and students participate in a digital environment based on new technologies),
- m-learning (methodology of teaching and learning that facilitates the construction of knowledge, problem-solving and the development of diverse skills and abilities in an autonomous and ubiquitous way, due to the mediation of portable mobile devices such as mobile phones or tablets),
- u-learning (ubiquitous learning based on learning environments that can be accessed in different contexts and situations, mainly via mobile devices),
- social learning (learning as a cognitive process that takes place in a social context and can occur purely through observation or by direct instruction),

- collaborative learning (an approach that tries to organize the activities in the classroom to become a social and academic learning experience, in which students work in groups to perform the tasks collectively), and
- Massive Open Online Course (MOOC, aimed at a large number of participants via Internet according to the principle of open and massive education).

The ODL platforms are mainly implemented as online web frameworks for education and therefore, they provide all the facilities for integrating the services used in the educational environment. These tools allow collaboration among all participants improving interaction. The increasing of the collaborative framework based on the web allows users to share information and take advantage of the interactions of other users. This fact leads to enhance their experience providing numerous benefits such as an increase in student motivation and the creation of collective intelligence [2]–[4]. This continuous development of web possibilities enhances the teaching-learning process and increases the effectiveness of learning systems in the knowledge society [5], [6]. However, there are still challenges that need to be addressed for effective ODL provision and to maintain excellent educational standards.

The online forum represents one of the most powerful and popular tools [7] and is frequently used communication technology tool in education. Online forums provide an excellent platform for learning and connecting students to the subject. They increase student engagement in the subject, promote deep learning, and maintain motivation [8]. However, the challenge they present is related to the task of managing the huge number of messages that are generated [9]. This can result in topics becoming fragmented over many threads with no search facilities to discover relevant information [8].

The review presented in this paper will describe the research into automated analysis of forum data. According to Hoogeveen *et al.* [10], forum research can be divided into two main groups: community question-answering (CQA) archives and discussion forums. Both groups promote community interaction and information sharing by the community. CQA archives are intended to assist people with problem solving and question answering. As soon as someone posts a good answer to a new question, the interaction is considered to be finished. Discussion forums on the other hand, are designed as a platform for discussion. However, the distinction between CQA archives and discussion forums is not very clear. Some discussion forums also focus on answering questions (for instance, Linux Questions –http://www. linuxquestions.org/ visited on 29th of November, 2018–) and specific CQA archives contain questions that are indeed conversations (for instance, Yahoo! Answers – https:// answers.yahoo.com/ visited on 29th of November, 2018–). Therefore, these two forum types share a number of characteristics which are not shared by other (semi) threaded discourses, like chat discussions, product reviews, or frequently asked question (FAQ) pages. Subsequently, these are outside the scope of this review.

With regard to the search and information management roles, they may be enhanced by technological advances and computing methods that facilitate the educational process. Thus, Artificial Intelligence (AI) approaches based on Natural Language Processing (NLP) can deliver tools specifically aimed at students and teachers. The objective would be to make the assimilation of content and course tracking easier for students. Furthermore, the tools would be designed to manage the learning platforms by teachers, especially in online courses –MOOCs and those based on collaborative learning [11]. In such courses, the forums are used both as a CQA platform and as a discussion forum.

The working hypothesis of our research is that this kind of solutions (AI approaches based on NLP) can be used to manage ODL learning platforms by teachers, as well as to manage the search mechanisms of both teachers and students. This will focus on online forums as they form the basis of collaborative learning, as the main tool for connecting students asynchronously. In addition, these forums allow better involvement of students and teachers in online platforms, allowing to improve knowledge management and teaching-learning process.

Our work addresses the main challenges related to ODL platforms and online forums, and the objectives are as follows:

I. Review the main approaches and techniques related to the management of ODL platforms and online courses.

II. Propose a set of analytics techniques (combining AI and NLP) and a general architecture that solve the main open issues found in the-state-of-the-art for managing online forums: (1) efficient tracking and monitoring of forums; (2) accurate information search; (3) extraction of relevant performance indicators.

The remainder of the paper is structured as follows. Section II summarizes the most relevant related work on forum analysis. Section III presents our proposal of the recommended techniques and the architecture for intelligent access to ODL platforms' forums and the extraction of relevant key performance indicators. Finally, the main contributions and our directions for future works are explained in Section IV.

II. REVIEW OF PREVIOUS WORK

This section summarizes the extensive work that has been done in this area and it has been organized into four subsections. The first one will deal with the work related to the management of discussion forums, explaining the different techniques of post classification and the main research areas of discussion forum analysis in educational environments. In the second subsection we will analyses the search within the forums of questions and answers (CQA archives) and how the answer retrieval can be improved by: classifying questions, subjectivity and user; post quality assessment; and, identifying similar questions. The third subsection will cover the previous work of NLP for the automatic analysis of massive amounts of posts, specifically on detection and

monitoring of topics, news and recommendation systems. We conclude this review with the fourth subsection that summarizes the findings after the analysis of previous work and the contributions of our proposal to overcome the main challenges and open issues in this area.

A. RELATED WORK ON DISCUSSION FORUM ANALYSIS

As previously mentioned, forums are a key tool in many online educational platforms. The advantages provided by this tool as well as some of its drawbacks are well known by the scientific community. In particular forums with a large number of users and a great volume of posts cannot be monitored effectively. It becomes very time consuming, and often impossible, to search for information, as students and teachers may not have the time to read all the posts generated by users on a weekly basis. This specific problem has been dealt with in many works, especially those related to online platforms directed towards large population groups such as generic forums (i.e., StackOverflow, Quora) or MOOCs [12].

Before tackling the problem of information search in discussion forums, we will present the post classification issue which can be used to improve the information retrieval process. In the work of Hoogeveen *et al.* [10], the authors emphasize that discussion forum threads are more dynamic than CQA ones: new questions can be asked in the middle of threads, topics can shift, and even though the initial post may be a question, it is by no means certain that the following posts contain answers. For this reason, it is very important the correct classification of forum posts. The authors distinguish four main types of post: (1) question posts, (2) answer posts, (3) acknowledgement posts, and (4) posts that contain the most important information of the thread (focus).

To identify question posts, classification techniques were used. Obasa *et al.* [13] used two types of features that complement each other: bag of word features together with simple rule features (the presence of question marks, and the presence of 5W1H words –what, why, when, who, where, how–) and with forum metadata features. In this task, the combinations of single features can achieve good performance, e.g. the authorship of the poster, the number of question marks, the number of 5W1H words and the number of posts in a thread [14].

To identify answer posts, the following are some examples of structural and content features that have been used: post author is not question author, the position of the post in the thread, whether the post is replied to by the question asker, whether a post contains a URL or not, etc. The best results were obtained by combining the two types of features [14]–[17]. Support Vector Machines (SVM) are the most commonly used models although some experiments using a semi-supervised co-training methodology have been carried out [16], [18]. The algorithm starts with a small number of training instances and continues for n iterations. At each iteration, two SVM classifiers are formed by training over two independent feature sets (structural features and pattern features). They are used to classify unlabeled instances. Finally, the predictions with the highest confidence are moved to the current set of labeled instances for training in the next iteration.

The same approach was used to identify acknowledgement posts. A positive acknowledgement post from the author of the question suggests that the problem is solved. On the contrary, a negative acknowledgement indicates that the proposed solutions are not correct. This information is very important for determining if an answer is useful or not.

In order to find the post with the thread focus Feng *et al.* [19] used an approach to detect conversation focus of threaded discussions by combining NLP analysis and IR (Information Retrieval) techniques. They took into account different features, such as lexical similarity, poster trustworthiness, and speech act relations in human conversations. They generated a weighted threaded discussion graph by applying feature-oriented link generation functions. Both quantitative and qualitative features were combined to analyses human conversations, specifically in the format of online discussions. The method presented in Mora *et al.* [20] also uses NLP techniques (Part-of-Speech tagging, partial parsing and semantic enrichment) to extract the relevant topics discussed in the online course forums. Furthermore, the method uses clustering techniques to obtain the different topics. A detailed explanation of this algorithm will be shown in Section III-C. It is important to mention that the information of the thread focus can potentially be used in thread summarization.

Discussion forum analysis in educational environments is divided into two main research areas: (a) studies on forum structure, user interactions, and types of student and teacher interventions; (b) studies focused on the content analysis of messages.

Concerning the first set of studies (a), there are works aimed at understanding how students and teachers are using the forum and how their learning and training expectations are met [21]. The procedures consist of comparing instructor and student participation rates, reviewing the role of the instructors, and analyzing the user interactivity to derive the learning outcomes and interaction patterns [8], [22], [23]. Suh and Lee [23] and Swan *et al.* [24] provide qualitative and quantitative measures to find key terms that make courses more attractive and better managed. Turning to the level of participation and interaction of the discussion group, encouraging student motivation is dealt with by Baxter and Haycock [25] and Yang *et al.* [26]. Performance is covered by Romero *et al.* [27], and indicators to predict the dropout rate of online courses and withdrawal from communities are presented by Yang *et al.* [26].

The second set of studies –focused on message content analysis (b)– intends to search information that assists users in meeting their learning or teaching objectives. These works focus on the forum's influence on student behavior and academic performance [28]. Most forum posts deal with reporting questions, errors, and discussion about course material and organization. According to several analysis schemes

provided by De Wever *et al.* [29], there is a wide variety of work related to content analysis of forum posts. The majority explore the effectiveness of online forums as platforms for innovating the educational practices of teachers [30], facilitating the teaching process [31], and providing indicators of student learning which can assist in student assessment [32]–[34]. This information provides teachers and forum administrators with an indication as to the development of critical thinking skills and whether learning has taken place. This information is very useful for managing online courses and designing learning strategies [35].

The methodology in both sets of studies previously described involves a variety of techniques. Firstly, there are studies that manually analyses forum content by labeling the posts by categories of interest [34], [36], [37]. Other studies use statistical research procedures for the post collected by means of questionnaires or surveys to the users [21], [32]. Secondly there are studies that use computational methods for automatic analysis of user messages that are sent to the forum. These computational methods perform interactions analysis on communication structures using a range of the following strategies which are subsequently described in more depth: data mining techniques [38]–[40]; social network analysis [22], [23], [26], [41]; and/or other artificial intelligence methods [42]–[44].

Data mining explores the posts extracted from forums to discover structures and to understand the dynamics of the community. There are several techniques that can automatically index, search, cluster, and structure the posts to discover a set of topics within a forum. These techniques are usually based on statistical topic models [38] and classification and clustering algorithms [27], [39].

Social network analysis offers a method for mapping group interaction, communication and dynamics. The method is usually implemented using computer-assisted qualitative data analysis software such as NVivo [45]. This method codes the contributions into units of meaning which are assigned to parts of messages based on semantic features.

Finally, other artificial intelligence methods for modeling dialogues and forum structures are based on machine learning approaches such as K-means clustering [46], Support Vector Machine [42], and Hidden Markov Models [43], [44].

B. RELATED WORK ON COMMUNITY QUESTION ANSWERING ANALYSIS

In this section we analyses the work directly related to one of the objectives of this paper, namely, including efficient search techniques in the CQA archives on different topics. An example of CQA archives is Stack Exchange. It is a network of question-and-answer websites on topics in diverse fields and is one of the most important CQA forums. The three most actively-viewed sites in this network are: Stack Overflow, Super User, and Ask Ubuntu. As mentioned in [47], it is not unusual to rely on such sources of information to find the correct answer to a given question. However, feeding forums with perpetual questions and answers makes this resource massive and full of duplicate posts and similar question variants. Thus, the search for an answer has become hard to achieve and led to the emergence of the area of research of CQA. The answer retrieval can be improved by classifying questions, subjectivity and user; post quality assessment; and, identifying similar questions. Next, we will explain these processes in detail.

Question classification is about detecting the type of question. There is no standard hierarchy or list of question types. Depending on the answer type that we would expect, the questions could be classified into: yes/no questions (the answer could potentially consist only of the word yes or no); opinion/topic (several different answers could be retrieved); and, factual questions (only one answer is the correct) [10]. Other researchers have instead used types that are closer to topics, based on the semantics of the question. They used question type taxonomies that are more fine-grained than taxonomies based on the question format or the answer types they are expected to receive [48]–[51].

We can distinguish three main strategies for automatic question classification in forum research community: (1) using the abovementioned question type taxonomies, (2) defining supervised and semi-supervised machine learning models with textual features [52], [53], and (3) specifying pattern matching systems using regular expressions [54]. Furthermore, automatic question type classification has been researched extensively outside of the forum domains, such as in the TREC (Text REtrieval Conference) Question Answering task [55]–[58] or in other environments [48], [50], [59]–[61].

Closely related to the problem of question classification is the subjectivity classification. It is an automatic process to determine whether or not the posts ask for or express an opinion. A good subjective answer should contain different viewpoints on the topic of the question, with arguments for and against. It is usually treated as a binary classification task, where questions are classified as being either subjective or objective. However, a third kind of question can sometimes be distinguished, namely, social questions [62]. Generally, three methods have been used to resolve the subjectivity classification problem: (1) specifying supervised [63]–[66] and semi-supervised models [18], [67], [68] in which different features were used: *n*-grams, question length, the time a question was posted, the topic of the subforum, punctuation marks, grammatical modifiers, etc.; (2) using both the question and its answers to construct the classification model [63], [69]; and, (3) defining a lexicon of words and multiword expressions, and a set of part-of-speech sequences with subjectivity weights manually constructed [70]. In the same way, in discussion forums the problem was treated as a classification task where complete threads were classified rather than individual posts. The main important research in this area is the work of Biyani [71] which uses the following features: structural, dialogue act, subjectivity lexicon-based, and sentiment.

With respect to the post quality it is important to emphasize that good access to high quality content has a high impact on user satisfaction and is the best way to retain existing users and attract new ones [10], [72]–[74]. We can distinguish two tasks in post quality research: (1) post quality classification into good/bad posts and (2) best answer identification. The techniques used in automatic post quality assessment were based on supervised classification models. They focused on feature engineering in both discussion forums [75]–[78] and CQA data [79]–[84]. In the process of best answer identification, the positive and negative votes posted by the users have been taken into account. Community generated answer scores or ratings are a good predictor of answer quality [79], [85]. Other techniques presented in the SemEval-2017 competition (explained below) classified the answers according to their relevance to the question from a question-answer thread [86].

Another factor that influences the post quality is the authorship of the posts. The users can be knowledgeable (experts) or not, good communicators or not, and willing to contribute quality content or not. High quality posts are often written by expert users. For this reason, user features are found to be helpful for post quality assessment [10]. Developing ways of identifying experts on forums (expert finding) can therefore help us to identify high quality content and vice versa [87].

Regarding the task of identifying expert users from less knowledgeable users we can distinguish four types of approaches [10]: (1) modeling the difficulty of questions—the expertise of the users is specified based on the difficulty of the question they have answered [88]—; (2) graph-based methods to identify expert users [89]–[97]. In these graphs, users are nodes, and edges are drawn from askers to answerers. The general idea is that users that ask high quality questions will receive many answers —they will have a high outdegree of edges—, and expert users tend to answer good questions —so they will have a high in-degree of edges— . These graphs can be enriched by representing questions and answers as nodes [98]. These approaches do not represent the ''possible'' interactions between the users based on their expertise, and therefore, researchers have completed the graph using, for example, user similarity [99]; (3) methods that use temporal information and the evolution of users to identify experts, future experts, or long-term contributors —who are usually experts—. They use time gaps between user postings or monitor the impact of changing time gaps over a specified time frame [100]–[102]; and, (4) methods based on deep learning. Wang *et al.* [103] use a Convolutional Neural Network (CNN) in which users are represented as vector representations of all the words in the questions to which they have given the best answer. At the end they are classified as expert and non-expert users.

In the CQA domain, the identification of similar questions is certainly an important preliminary step for providing a correct answer to a posted question. It is necessary to figure out if a question has not already been treated in other posts, essentially for a matter of response effectiveness and to reduce as much as possible duplicate posts. To that end, question-to-question similarity task offers a key challenge while it has to deal not only with similar questions in terms of lexical similarity but also in terms of reformulation, paraphrasing, semantics, etc.

The CQA forums are increasingly gaining popularity. Consequently, since 2015 during a series of ongoing annual competitions SemEval –Semantic Evaluation– (http://en.wikipedia.org/wiki/SemEval visited on 29th of November, 2018) one of the tasks that has been developed is denominated Track 3 (http://alt.qcri.org/semeval2017/task3/). Track 3 evaluates systems that carry out the automatic process of finding good answers to new questions in a forum of discussion created by the community. For example, by retrieving similar questions in the forum, the correct answers can be identified.

In the SemEval-2017 competition, two additional subtasks to Track 3 have been proposed. In the first subtask, given a new question and a set of related questions from the collection, similar questions needed to be classified according to their similarity to the original new question (with the idea that the answers to similar questions must answer the original question also). In the second subtask, given a question from a question-answer thread, all the response posts were classified according to their relevance to the question. A detailed description of the Semeval 2016 and 2017 editions and their participants is presented in Nakov *et al.* [86] and Nakov *et al.* [104].

These CQA forums are rarely moderated. They are generally open tools, and therefore, there are few restrictions, if any, about who can post and who can answer a question. The main advantage of this is that anyone can freely ask questions and generally expect correct and honest answers. By contrast, an important disadvantage is that a major effort is required to analyses all the answers and make sense of them. For example, a question usually has hundreds of answers which makes it very time consuming for the user to analyses them. The main problem that needs to be addressed by researchers is that there is a lot of irrelevant material, given that online forums are a resource created by a community of occasional users. Furthermore, informal language is used and there is often a large number of spelling and grammatical mistakes.

The following techniques were used in this area: neural networks which use deep learning methods –Recurrent Neural Networks, RNN–; Long Short-Term Memory –LSTM– to capture long distance dependencies; Convolutional Neural Networks –CNN– [105], [106]; and SVM [107]. Although these approaches demonstrate remarkable improvements in a wide range of applications, their computational cost and the need for an extensive training data make them inefficient with small and specific datasets.

The following discourse based techniques were adopted: word-based methods [108]–[111]; more complex discursive structures such as phrases, sentences, paragraphs, or documents [106], [110], [112]–[115]. Furthermore, the approach

of Hazem *et al.* [47] was used in which each question is represented by the element-wise addition of its words embedding vectors and they are represented in a joint sub-space where similar pairs are moved closer thanks to a mapping matrix.

C. RELATED WORK ON TOPIC DETECTION AND TRACKING, RECOMMENDATION AND NEWS TRACKERS SYSTEMS

The use of NLP as a method for automatic analysis of mass quantities of texts has been widely studied. Studies have been done on analysis of social networks and other online platforms, however, these techniques have not been generally used in the context of educational courses with massive forums. Consequently, these methods are potentially useful for improving the learning-teaching process, but are still relatively immature for educational applications.

With regard to the automatic analysis of text, several research lines are being currently developed: Topic Detection and Tracking, Recommender or Recommendation systems, and News Trackers systems. These lines are detailed next.

Topic Detection and Tracking (TDT) was developed by Defense Advanced Research Projects Agency (DARPA) to assist the detecting and following of new events in a stream of broadcast news stories as well as their reappearance and evolution [116]. TDT techniques are applied to Social Networks (e.g., in [117]) in real data sets. The TDT cluster detection technology was deployed in a real world setting, and several drawbacks were solved regarding the incremental clustering over time, which is a key issue in forum analysis [118].

Recommender or Recommendation Systems (RS) help to determine which information should be offered to individual consumers and allow users to quickly find the personalized information that suits their needs [119]. RSs are presently ubiquitous in various domains and e-commerce platforms, including book recommendations at Amazon, music at Last.fm, movies at Netflix and references at CiteULike.

Similarly, Collaborative filtering (CF) approaches have been extensively investigated in the research community and have a wide application in industry. They are based on a rather simplistic assumption that if users ranked items similarly in the past, then they are likely replicate the same ranking in the future [120]. Winoto *et al.* [121] make personalized paper recommendations for users in the education sector. For instance, when the RS guides the tutor or student in the selections of relevant courses, programs, or learning materials (books, articles, exams, etc.), and the selection criteria enables the user's learning goals, background knowledge, motivation, among other things to be included. In the same way, Sathick and Venkat [122] facilitate the career guidance of online learners, who pursue their graduation in an open and distance learning environment, by implementing an online recommender application. Their objective is to facilitate user acquisition of semantic knowledge from heterogeneous web sources and decision making.

Finally, the News Trackers systems (NT) apply TDT techniques in the real world. For instance, the MemeTracker (http://www.memetracker.org/ visited on 29th of November, 2018) by Leskovec *et al.* [123] constructs daily news cycle maps by analyzing approximately 900,000 news stories and blog posts per day from 1 million online sources that range from mass media to personal blogs. All quotes and phrases that appear most frequently over time are tracked. The set of quoted phrases and sentences found in the articles will act as tracers for memes. This makes it possible to see how different topics are reported on a daily basis in the news and blogs as well as how certain stories persist while others fade. Currently, many News Tracker applications are being developed, mainly in online newspapers such as NewsTracker (http://www.yournewstracker.com/ visited on 29th of November, 2018). In this field, the analysis of microblog service providers (such as Twitter) are using lexical matching [124] and semantic features [125] to categorize posts and to join particular topics.

D. FINDINGS AND CONTRIBUTIONS OF THE PROPOSAL

After reviewing the previous work (summarized in Table 1), next we present the main challenges and open issues in this area:

- An efficient tracking and monitoring of forums is required. It is necessary to manage the high number of messages generated in forums and take into account their characteristics. These messages are unordered, unstructured and cover a great variety of topics.
- Different search tasks on the forums are needed: (1) the most similar question to the original question, and (2) the most likely answer. It is very important to determine if a question is new (it does not exist in the forum) to find the correct answer. Otherwise, if the question is similar to an existing question, then the most likely answer should be found.
- Performance indicators must be extracted from forums with the aim of managing online courses and designing efficient learning strategies.

Table 1 indicates three characteristics of the different proposals: (1) the references of the previous works; (2) the classification of the works; and (3) the functionalities and techniques used.

Next, we summarize the main contributions presented in this paper:

- An exhaustive review of the approaches and techniques related to the management of online courses.
- Proposal of a set of analytics techniques and a general architecture that solve the main challenges found in the-state-of-the-art to manage online forums more efficiently:

(1) Efficient tracking and monitoring of forums generated.
(2) Accurate information search to find: (i) the most similar question, and (ii) the most likely answer.

TABLE 1. Summary of previous work.

References	Classification	Functionalities and techniques
[13]—[20]	Discussion forum analysis	Post classification: question posts, answer posts, acknowledgement posts, posts that contain the most important information of the thread (focus). Techniques: SVM, semi-supervised AI methods, NLP + IR methods.
[71]		Subjectivity classification: subjective questions, objective questions. Techniques: supervised machine learning models.
[75]—[78]		Post quality classification. Techniques: supervised models.
[8], [21]—[27]		Forum structure: Analysis of teacher and student participation rates. Extraction of qualitative and quantitative measures (student and performance motivation, dropout rate, etc.).
[28]—[35]		Content analysis of messages: find information that assists users in meeting their learning or teaching objectives. Understand the effectiveness of forums for innovating the educational practices of teachers, facilitating the teaching process and providing indicators of student learning.
[21], [26], [27], [32], [34], [36]—[46]		Techniques: manual analysis. Statistical procedures. Computational methods based on Data Mining, social network analysis and AI methods (SVM, HMM, etc.).
[48]—[61]	CQA, Community Question Answering analysis	Question classification: yes/no questions, opinion/topic, factual questions. Techniques: taxonomies, supervised and semi-supervised machine learning models with textual features, pattern matching.
[18], [63]—[70]		Subjectivity classification: subjective questions, objective questions. Techniques: supervised and semi-supervised models, using both the question and its answers to construct the classification model, defining a lexicon with subjectivity weights.
[79]—[86]		Post quality research: post quality classification, best answer identification. Techniques: supervised models.
[88]—[103]		Expert finding and user classification. Techniques: modelling the difficulty of questions, Graph-based methods, methods that use temporal information and the evolution of users, deep learning.
[86], [104]		Answer search to new questions in online forums created by the community. Identification of the correct answers. Classification of similar questions. Classification of answer posts according to their relevance with respect to the original question.
[47], [105]—[115]		Techniques: RNN, LSTM, CNN, SVM. Word-based methods. Methods based on complex discursive structures: phrases, sentences, paragraphs, and documents.
[116]—[118]	TDT, Topic Detection and Tracking	Techniques for detecting the appearance of new topics and for tracking the reappearance and the evolution of them. Application of TDT techniques to social networks.
[123]	NT, News Trackers systems	Applications of TDT techniques in the real world. Maps of the daily news cycle.
[119]—[122]	RS, Recommendation Systems	Analysis of information that has to be offered to consumers to find the personalized information that fits their needs. Collaborative Filtering techniques. Recommendations to the teacher/student to pick relevant courses, programs, or learning materials (books, articles, exams, etc.).

(3) Extraction of relevant performance indicators with the use of Data Mining (DM) techniques.

III. THE PROPOSED ANALYTICS TECHNIQUES AND ARCHITECTURE

As previously mentioned in Section I the working hypothesis of our research is that Artificial Intelligence approaches based on Natural Language Processing can be used to manage ODL learning platforms by teachers and administrators. We are focusing on online forums because they are essential for collaborative learning as main tool for connecting students asynchronously.

The studies carried out on this topic reveal that online forums have a high impact on how knowledge is transferred among students [20], [25]. These findings can be exploited to improve student engagement, retention and learning. However, no effective solutions have been proposed to facilitate the automatic monitoring of online forums and turn them

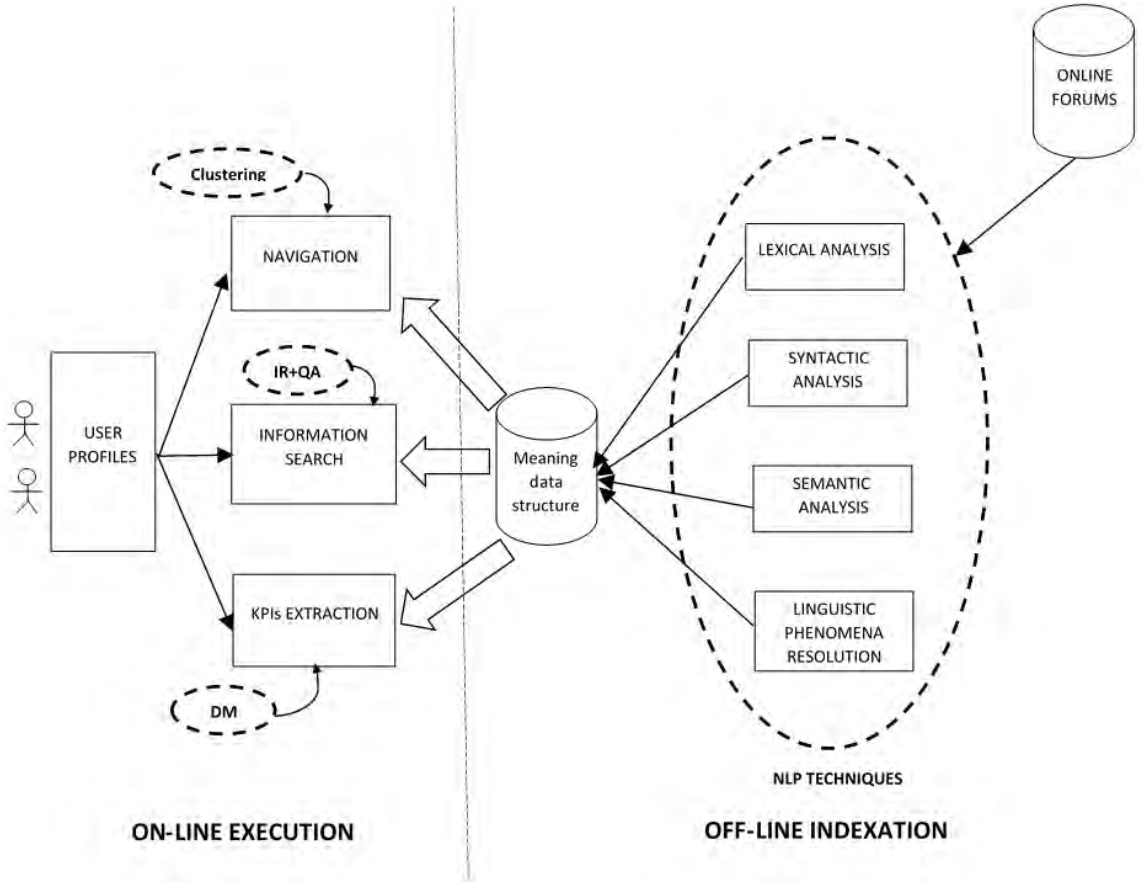

FIGURE 1. General architecture and analytic techniques for an efficient management of forums.

in a good tool for interaction and communication, specially even when excessive messaging occurs in a disorderly and unstructured manner.

In this context, to find the appropriate answer in the forums (question and answer searches in CQA) and when to create a new thread are important issues. As mentioned in Section II, there are a set of challenges for taking advantage of the potential of such tools: (a) existence of irrelevant material; (b) the data contain a lot of noise, therefore, it is difficult to retrieve relevant material; (c) informal use of language; (d) spelling and grammatical errors. In general, existing proposals require a high computational cost and a large amount of data. In addition, they are inefficient in small and specific data sets.

Finally, a more important issue to the proper use of forums for educational purposes is the extraction of Key Performance Indicators (KPIs) that calculate, among others, the participation, motivation, dropout rates and performance of students. Furthermore, enriched information can be retrieved from forums. For example, the questions asked by the students can be grouped by similarity and thus to detect possible problems in the students' learning. As presented in Section II, this information helps the teachers and administrators of the forum to know if students are developing critical thinking skills and if learning actually occurs. This information is very useful for managing online courses and designing learning strategies.

Under this approach, the main aim of this work is to propose a set of learning analytics techniques based on NLP to overcome the existing challenges. These techniques will be used to design automatic analysis tools to facilitate the monitoring of online learning communities.

The NLP techniques that we considered crucial to achieve our aims are: lexical, syntactic and semantic analysis, ellipsis and anaphora resolution, Information Retrieval (IR), Question Answering (QA) and clustering. Lastly, DM procedures will infer the KPIs.

In Figure 1 the proposed general architecture that makes use of the NLP techniques is presented. The techniques require the coordination of different research areas. Each research area presents an extensive previous work, with a plethora of off-the-shelf tools, which forces the architecture to be modular, in order to facilitate the exchange between different tools. The linking between the lexical, syntactic, semantic analysis, and linguistic phenomena resolution will be carried out by means of the meaning data structure detailed in subsection III-B, which will compile all the knowledge provided by the different tools, in a unique and unambiguous

way. The generation of this structure, as well as the input and output of each module in the architecture —i.e. the navigation, information search and KPIs extraction— will be accomplished by a set of interface modules. The main aim of these interfaces is to map the expected input/output of each tool into the expected format in the architecture, which will be in XML format, in order to make easier the selection of different tools for the same task. For example, our architecture has been evaluated with the FreeLing POS tagger tool, which generates its output as it is presented in Figure 3. In case the TreeTagger tool is used instead of FreeLing, we only have to update the corresponding interface to generate the XML meaning structure, with the proper mapping between their lexical tags (e.g. the ''Fp'' tag will be mapped with the ''SENT'' tag in TreeTagger).

Moreover, the proposed architecture in Figure 1 presents two phases, off-line and on-line. On the right side, the off-line phase of processing the forums is carried out before the online execution. The forums are pre-processed using the NLP techniques. After that, the process of indexing is performed —similar to IR or QA systems— in which the information of the collected forums is organized to access them more easily in the information search process. Given the highly dynamic nature of online forums, the indexation techniques must be robust especially for incremental and update processes. The output of this stage is the meaning data structure with lexical, syntactic and semantic information that unambiguously represents the meaning of the text in XML format (see subsection III-B). It will be used for the communication between the different modules of the architecture. On the left side of the figure, the on-line process related to the interaction with the user is shown. The three main functionalities of the system —navigation, information search and KPI extraction— , with the involved techniques —clustering; Information Retrieval + Question Answering; Data Mining— are specified.

For the purpose of checking the advantages of our proposal (see Section II-D), we have carried out a preliminary evaluation of each key feature and technique proposed in our architecture, by selecting a set of up-to-date and competitive tools that provide the functionalities required in our proposal. Each of these tools have been properly evaluated and compared with related work in their specific research areas, which assures the effectiveness of the architecture presented in this paper. Moreover, we should stress that the modular design of our architecture facilitates the use of alternative tools. This preliminary evaluation has been performed on a case study, which has been sourced from Stanford University's Statistics in Medicine online course, MEDSTATS (https://lagunita.stanford.edu/courses/Medicine/MedStats./Summer2015/about visited on 29th of November, 2018). MEDSTATS course aims to provide a firm grounding in the foundations of probability and statistics. Specific topics include: 1. Describing data; 2. Statistical inference; 3. Specific statistical tests. The course focuses on real examples from both medical literature and mass media sources, and is divided into nine units, described in detail in the course program (shown in Appendix Section).

TABLE 2. Forum categories in *Statistics in Medicine (MEDSTATS)* online course.

Forum Categories	Threads	Posts
Course Material Feedback	178	562
External Resources	9	29
General	58	167
Homework	78	316
Introductions	1	1
Platform Feedback	1	4
Study Groups	1	3
Tech Support	13	56
Videos	171	488
Unit 1 (modules 1, 2, etc.), Unit 2…Unit 9 Sub-forums	171	479
TOTAL	**681**	**2,105**

Each course unit has an associated forum. In addition, the system administrator, according to the recommendations of the teachers, has defined different forum categories related to topics of interest. When a student creates a new thread, he or she decides to associate the thread to a forum category. Furthermore, both the type of the intervention —discussion, question or answer—as well as the person who adds the post are specified. Two general types of profiles are definable depending on the author of the post: ''student'' and ''teaching assistant''. During the course analyzed, 2,105 posts were written by the students and the teaching assistants. This is an indication of the volume of information that must be processed by teachers and students. In this particular course there were 9 general forums and 96 sub-forums of the 9 Units modules. In Table 2 a summarization of the categories, the number of threads and posts is detailed.

Our case study focuses on the ''Course Material Feedback'' forum category (which contains 178 threads and 562 posts). We have chosen this category because it contains the largest number of threads and posts and it has interventions about diverse issues.

The following subsections (from III-A to III-E) describe the key features and techniques proposed in our architecture (Figure 1) showing its benefits on its application to the previously described case study. In first place, the user profiles management are introduced; next, the meaning data structure involved on analyzing and messaging is described; and the final three sections report the modules that allow to carry out the mentioned three functionalities that solve the main deficiencies found in previous work.

A. USER PROFILES MANAGEMENT

In the first place, it is necessary to identify the requirements of the users of forums generated in ODL learning platforms for management, administration and search tasks. This information is required both for teachers and students.

Each user of these forums presents different needs so that each type of user must be adequately defined through profile management. For example, the relevance of each post

can be assessed with a different weight depending on the parameters that are determined for each user. This idea is already implemented in social networks or in the search on the web. In this way, different levels will be specified for the student profile. Similarly, a profile will be specified for teachers by distinguishing different types according to their responsibility (e.g. director or teacher of a course). For this task (user classification or expert finding) the approaches presented in Section II-B could be used.

The requirements of each group will be deducted from their particular objectives in learning and teaching process:

- The objectives of the teachers should be focused on the monitoring of the operation of the course, so they will require tools that allow them to verify that the contents are being assimilated properly, that have not arisen possible problems in the development of the course, etc.
- The objectives of students should be focused on learning, that is, on getting tools to help them assimilate content, either by locating the source offered by teachers to help solve problems, or those offered by the classmates of the course, based on questions asked and answered previously.

To complete the user requirements and their scenarios, the discursive and linguistic structure of the online forums will be analyzed. This process takes into account that they can be of different types and domains. The forum characterization will produce information on number of threads and posts, number of participants, types of linguistic problems to solve (anaphora, ellipsis, etc.), language model used in the forum, differentiating characteristics between different domains of forums, necessary semantic resources, etc. The rhetoric figures such as anaphora or ellipsis should be resolved to get an adequate understanding of the meaning of the text, as well as to make a lexical, syntactic or semantic analysis in a more precise way. Poor contexts are also expected, with short phrases, as well as spelling errors, abbreviations and informal language, which will cause problems in ambiguity resolution tasks.

B. MEANING DATA STRUCTURE

After analyzing the forums in a context of dialogue, a structure will be defined for the representation of the meaning of the text. This information must be described in a unique and unambiguous way, to achieve the desired precision and should serve as an element of communication between the modules, resources and techniques that make up the system.

An example of this type of meaning data structure that uses NLP techniques is presented in the work of Martínez-Barco *et al.* [126], which the authors call L-Bricks. In their work, the authors present a unique model for the understanding and the generation of the Human Language based on techniques of deconstruction of the language. In its model, we can emphasize two layers: (1) Understanding Layer of the human language, and (2) Deconstruction Layer of language in basic units of knowledge.

(1) Understanding Layer of Human Language: in this layer all the resources/tools (lexical, syntactic and semantic) and necessary techniques (such as the resolution of linguistic problems –anaphoric expressions and ellipsis– and the resolution of temporal references –identifying the events and the time in which they occur–) are collected, analyzed and integrated in order to transform the information obtained from different sources (in our case, textual information obtained from the forums) into useful knowledge. This will later be stored in the basic knowledge units.

(2) Deconstruction Layer of the language in basic units of knowledge: in this layer the activities related to the definition, structuring and insertion of the data previously obtained in these basic units are processed. The information units are called L-Bricks (Language Brick). The authors define 3 properties for L-Bricks: a) BRICKS.DIMS (Dimensions), in which will define the multidimensional structure of the brick; b) LBRICKS.OPRS (Operations), whose objective is the planning of the set of possible operations for the L-Brick unit, and; c) LBRICKS.IMPL (Implementation), whose purpose is the computational implementation of L-Brick structure, operations and storage.

Next, the application of our architecture on the case study is shown. The off-line indexation process input is the ''Course Material Feedback forum''. In this way, we can analyses different sets of threads in a flexible manner. For example, we can study the threads of the same forum category, such as ''External Resources'', or the threads between different dates (i.e., from the beginning up to mid semester). It is also possible to analyses all the threads of two or more forum categories. Each post is processed individually. The output of the off-line process is the post which contains lexical, morphological, syntactic, and semantic information. It is important to emphasize that linguistic phenomena resolution (such as definite description or anaphora resolution) is carried out.

Below, we will show the application of the previous mentioned stages with a real example extracted from the ''Course Material Feedback'' forum. In Figure 2 thread#38 and thread#142 are shown. Each post begins with the type of the intervention (discussion/question/answer), the person who adds the post and his/her profile (student/Teaching Assistant). For instance the second post of the thread#142 begins with the following specification: *<type="answer"; name="JWallach"; profile="Teaching Assistant ">*. Following this, the text of the post is specified.

Let us assume that we are processing the second post of the thread#142. All the previous threads (with all their posts) have been processed. The stage of the lexical-morphological analysis consists of carrying out the PoS (Part-of-Speech) tagging of the text. The output is a set of pairs <word, PoS tag> where the PoS tag identifies the grammatical category (noun, verb, adjective, adverb, conjunction, etc.) and the morphological information (singular, plural, masculine, feminine, etc.) of the word. In Figure 3 the output of the lexical analysis stage is shown. For example, the word *factors*

THREAD #38
<type="question"; name="anonymous"; profile="student">
forward and backward regression
How to do forward and backwards regression are they different

<type="answer"; name="JWallach"; profile="Teaching Assistant ">
These both have to do with variable selection - determining which of your variables should be included in a model. Forward and backward regression are two step-wise regressions procedures.

You can think of backwards selection and backwards elimination. …

With forward selection, you just reverse the backwards method!. …

Hope this helps! Josh

<type="answer"; name="JoeDoe2014"; profile="student">
https://en.wikipedia.org/wiki/Stepwise_regression

<type="answer"; name="nivinsharawy"; profile="student">
Thanks one more question in this subject If just do multiple linear regression is it efficient?

THREAD #142
<type="question"; name="Ashraf91"; profile="student">
adjusted and unadjusted risk ratios
Hi all ,

I would like to understand what is the difference between the adjusted and unadjusted risk ratio and how to calculate them ?

Best , Ashraf.

<type="answer"; name="JWallach"; profile="Teaching Assistant ">
Hello Ashraf,

You will learn about this more in the later units! This has to do with regression analyses when you include multiple factors/variables. In particular, a logistic regression can be used to estimate adjusted odds ratios and these can usually be interpreted as risk ratios (when outcomes are rare). They are adjusted because by including other predictors, you have "adjusted out" or "controlled out" the influence of other factors.

best, Josh

FIGURE 2. **Examples of threads extracted from the "Course Material Feedback" forum.**

Hello **NN** Ashraf **NPS** , **Fc** You **PRP** will **MD** learn **VB** about **IN** this **DT** more **RB** in **IN** the **DT** later **RBR** units **NNS** ! **Fat** This **DT** has **VBZ** to **TO** do **VB** with **IN** regression **NN** analyses **NNS** when **WRB** you **PRP** include **VBP** multiple **JJ** factors **NNS** / **Fh** variables **NNS** . **Fp**
. . . .

FIGURE 3. **Example of lexical-morphological analysis.**

is tagged as **NNS** (noun plural) whereas *Ashraf* is tagged as **NPS** (proper noun singular). We have used the FreeLing POS tagger [127] which uses a tagset based on Penn TreeBank tagset.

The syntactic analysis stage consists of performing the parsing of the text. The text of the post is partially parsed to extract noun phrases (NP), prepositional phrases (PP), and verbal phrases (VP). The text chunks not included in these phrases are skipped (SK) in the parsing. By contrast, NPs

***np*(Hello Ashraf)**, sk(,),
***np*(You)**, ***vp*(will learn)**, sk(about), sk(this), sk(more), ***pp*(in *np*(the later units))**, sk(!)
sk(This), ***vp*(has to do)**, ***pp*(with *np*(regression analyses))**, sk(when), ***np*(you)**, ***vp*(include)**, ***np*(multiple factors)**, sk(/), ***np*(variables)**, sk(.)
sk(In_particular), sk(,) ***np*(a logistic regression)**, ***vp*(can be used to estimate)**, ***np*(adjusted odd ratios)**, sk(and), ***np*(these)**, ***vp*(can)**, sk(usually), ***vp*(be interpreted)**, ***pp*(as *np*(risk ratios))**, sk((), sk(when), ***np*(outcomes)**, ***vp*(are)**, sk(rare), sk()), sk(.)
***np*(They)**, ***vp*(are adjusted)**, sk(by), ***vp*(including)**, ***np*(other predictors)**, sk(,), ***np*(you)**, ***vp*(have)**, sk("), ***vp*(adjusted out)**, sk("), sk(or), sk("), ***vp*(controlled out)**, sk("), ***np*(the influence *pp*(of *np*(other factors)))**, sk(.)
. . . .

FIGURE 4. **Example of syntactic analysis.**

np(a logistic **({02997650} <adj.pert>)** regression**({06036794} <noun.cognition>)**)
pp(with *np*(regression**({06036794} <noun.cognition>)**
analyses**({06023392} <noun.cognition>)**))
vp(include**({02639021} <verb.stative>)**)
. . . .

FIGURE 5. **Example of semantic analysis.**

can have nested structures such as PPs, appositions or relative clauses; therefore, these phrases are fully parsed. Moreover, coordinated NPs and PPs are parsed. These phrases represent the "main concepts" involved in the text. We have used the partial parser presented in the work of Ferrández *et al.* [128]. In Figure 4 the syntactic analysis of the abovementioned post is shown. We can observe the different phrases that have been partially parsed and the skipped phrases. For instance, the noun phrase *np(a logistic regression)*, the prepositional phrase *pp(with np(regression analyses))* which includes a nested noun phrase, or the verbal phrase *vp(include)*.

The objective of the semantic analysis stage is to enrich the post information with semantic information. It is obtained from additional semantic resources, such as WordNet (http://wordnetweb.princeton.edu/perl/webwn visited on 29th of November, 2018) and the course program ontology. In this way, semantic comparisons can be done such as synonymy or hyponymy. In Figure 5, the semantic information of different phrases is shown. Each noun, verb, adjective and adverb is labeled with its synset (number which identifies a set of synonyms) and its type. For example *regression* has the synset number *06036794* and belongs to the type *noun.cognition*.

The final stage of the off-line indexation process is the linguistic phenomena resolution (such as definite description or anaphora resolution). The objective of this stage is to resolve the referential ambiguity of the text. To do this, the anaphoric expressions are resolved and replaced by the entities to which they refer. We have followed the algorithm presented in Palomar *et al.* [129]. It is important to highlight that by resolving the linguistic phenomena we improve the

comprehension and coherence of the text as the anaphoric elements are replaced by the entities/concepts to which they refer. In our example (the second post of the thread#142) two pronominal anaphors are resolved: the pronouns *these* and *they*. Both anaphoric expressions are replaced by the noun phrase *adjusted odds ratios*.

The output of the off-line process is the meaning data structure with lexical, syntactic and semantic information that unambiguously (linguistic phenomena have been resolved) represents the meaning of the text. In our evaluation we have used a structure (specifically a list), that stores the main concepts in the sequential order as they appear in the text. This list contains a set of structures generated by the partial parser [128] called *slot structures* (SS). They store an identifier (marked as upper cases such as X), the morphological knowledge (in the structure ''conc,'' such as number and gender), syntactic knowledge (e.g., the slot structures of nested phrases), semantic information (the synset and the type), the term as it appears in the text (e.g., *used*), and the lemma (e.g., *use*).

C. BROWSING AND AUTOMATIC MONITORING OF FORUMS

The first functionality of the system consists in being able to browse automatically the structure of the forum generated by the application.

A computational method to facilitate the tracking and monitoring of forums generated by online learning courses and communities will be used for the abovementioned purpose. For instance, the method presented in Mora *et al.* [20] could be used. It analyses the forum information through Natural Language Processing techniques and extract the main topics discussed in the forums according to the subject matter being studied in order to describe its content and evolution along the course. In detail, the text corresponding to the program of the subject is processed by NLP tools (Part-of-Speech tagging, partial parsing and semantic enrichment). After that, it is clustered and the most relevant topics of each cluster are extracted. Next, each new student post is processed in a similar way in order to obtain the clusters of the set formed by the original post and the replies to it. In this way, in each separated thread of posts, they accomplish the anaphora resolution of definite descriptions which are included in the same cluster. Finally, the list of the most relevant topics from all the posts is processed jointly with the subject topics in order to obtain the final list of relevant topics linked to the subject program topics.

Based on the extracted information, many actions can be done: for example, the system can automatically restructure the forum categories according the main themes addressed, the teacher staff can identify what the hot issues are in order to provide useful explanations about them, and the students can identify easily the thread where to look for their answer or to post new comments. In this way, this information performs tracking the forums and the online course more effectively both for students and instructors.

[
1 [np*(Multiple Linear Regression and *np*(Statistical Adjustment)), *np*(Categorical Predictors *pp in* np(Regression)), *np*(Simple Linear Regression), *np*(Linear Regression Results), *np*(Regression Analysis), *np*(Regression Worries)),
*2 [np(*Covariance and *np(*Correlation*))],*
*3 [np(*Residual Analysis*)],*
*4 [np(*Practice Interpreting*)],*
*5 [np(*Overfitting and *np(*Missing Data*))],*
. . . .

FIGURE 6. Cluster extraction of course program. Unit 8: Regression analysis.

To evaluate this functionality, we have applied the abovementioned method proposed in Mora *et al.* [20]. First, the text corresponding to the course program was processed by the previous mentioned NLP tools (Part-of-Speech tagging, partial parsing, semantic enrichment and linguistic phenomena resolution) obtaining the mentioned meaning data structure. This structure was the input for the clustering process. After that, the similarity matrix between each pair of parsed phrases was calculated obtaining a list of clusters. Subsequently, clusters were ranked according to their relevance using Information Retrieval techniques. Finally, the most relevant topics of each cluster according to their relevance were extracted.

For instance, related to our example, after processing unit 8 in the course program, the five clusters shown in Figure 6 are identified.

Following this, each new student post was processed in a similar way to obtain the clusters and the relevant topics of the set formed by the original post and the subsequent replies.

We processed the second post of the thread#142 (Figure 2). All the previous posts had been processed and the new terms were written in bold. To conclude, a filtering process was applied to the set of clusters so as to filter some frequent and irrelevant expressions (e.g., ''Hi'', ''are,'' or ''the difference''). In Figure 7, an excerpt of the extracted clusters is shown.

Finally, the list of the most relevant topics from all the posts was processed jointly with the course program topics to obtain the final list of relevant topics linked to the course program. For example, after merging cluster 1 of Figure 6 (clusters of course program) and cluster 1 of Figure 7 (clusters of student posts) a cluster with the topic of regression was obtained. The final result is a new topic (regression), which was added to the list of relevant topics.

In Table 3, the main topics of the ''Course Material Feedback'' forum category that were detected by the system are shown. We can distinguish 14 main topics ranked by the number of threads related to each topic. For instance, the most important topic is ''Homework Questions'' with a percentage of 46.1% of the total threads in this forum category. By contrast, ''Califications'' is the least relevant topic with 1.1% of the threads.

As previously discussed, one of the main benefits of this work is the automatic processing of a huge amount of posts

1 [[np(forward and *np*(backward regression), *np*(forward and *np*(backwards regression), *np*(two step-wise regressions procedures), *np*(backwards method), *np*(multiple linear regression), ***pp*(with *np*(regression analyses)), *np*(a logistic regression)]**,
2 [vp(to do), ***vp*(have to do)**, *vp*(do)*]*,
3 [vp(are)*]*,
*4 [pp(*with *np(*variable selection*)), pp(*of *np*(your variables)), *pp(*of *np(*backwards selection and *np*(backwards elimination))), *pp*(with *np*(forward selection)), ***np*(multiple factors/variables)]**,
*5 [vp(*determining)*]*,
6 [vp(should be included)*]*,
*7 [pp(*in *np(*a model))*]*,
8 [vp(can think)*]*,
9 [vp(reverse)*]*,
10 [vp(hope)*]*,
11 [vp(helps)*]*,
12 [np(Josh)*]*,
13 [np(https://en.wikipedia.org/wiki/Stepwise_regression)*]*,
14 [np(one more question)*]*,
*15 [pp(*in *np*(this subject))*]*,
16 [np(adjusted and unadjusted risk ratios)*]*,
17 [np(Hi), *np*(Hello)*]*,
18 [vp(would like)*]*,
19 [vp(to understand)*]*,
20 [np(the difference)*]*,
21 [vp(to calculate)*]*,
*22 [**np*(Ashraf)**]**,
23 [vp*(will learn)*],
24 [pp*(in *np*(the later units))*],
25 [vp(*include*)],
. . . .

FIGURE 7. Cluster extraction of student posts.

TABLE 3. Main topics of the "Course Material Feedback" forum category extracted by the system.

Topics	%Threads
Homework Questions	46.1%
Congratulations	11.8%
homework submission Problems	6.7%
video quality, transcribed notes, sound / problems	6.3%
Optional/ additional Material	5.6%
SPSS & Deducer & R	4.5%
Join/ start now?	3.9%
Regression	3.4%
Certificate	2.8%
Correlation	2.2%
Homework due date?	2.2%
Final Exam	1.7%
Mc Nemar test	1.7%
Califications	1.1%

included in the forums. The results of this study show the importance of our proposal because it is possible to display the main topics in a concise and classified manner. Based on the extracted information, many actions can be carried out: (1) the system can automatically restructure the forum categories according to the main topics being addressed; (2) the instructors can identify what are the trending issues providing feedback; and (3) the students can easily identify the thread to look for answers to their questions, or to post new comments. In summary, information about main topics allows students and instructors to track more effectively the forums and the online course. In this way, we achieve a comprehensive solution and maximize the benefits from the information generated in these forums, and we overcome partial solutions proposed in previous work (e.g. the one in Wise *et al.* [130], in which the authors only classify threads that are substantially related or unrelated to the course material, by means of learning linguistic features of each category). The advantage of our approach is the combination of the forum topics with the course ones, and, in addition, our architecture includes the improved services, which are explained in the next two subsections (information search and KPIs extraction).

D. INFORMATION SEARCH FOR QUESTIONS AND ANSWERS

This functionality will offer the traditional format of a general-purpose search engine (QA system), in which the user makes a request for information and the system will present the most relevant results, either the most similar question or the answer more likely to the question.

Unlike traditional search engines, the starting point of a QA system is not to return a list of documents, but to provide a list of "pieces of text" that supposedly contain the information required by the user. To achieve this goal, a greater understanding of documents will be required, which will cause delays to the search process, implying longer response times. However, this inconvenience is currently addressed by reducing the size of the text on which more expensive computational techniques such as NLP will be applied. Usually, the work is done on the output of a traditional IR system. That is to say, only a limited number (usually hundreds of documents) from the millions of available documents is selected to perform the costly techniques.

Another aspect that characterizes the QA systems is that the question needs to be expressed in its complete form in natural language, that is, in the highest level of detail possible. In contrast, traditional search engines only require as input a sequence of keywords. Examples of NLP-based information search techniques are the works by Ferrández [131] and Muñoz-Terol *et al.* [132]. Other sets of techniques used in QA have been presented in the TREC (Text REtrieval Conference) competitions since 1999 (http://trec.nist.gov/ visited on 29th of November, 2018).

Recently, specific methods have been developed for information search in the forums [10] which obtain better results than the proposals presented in the competitions SemEval (2016 and 2017) in "question-to-question similarity" task [47]. The evaluation accomplished in this paper has run the AliQAn system [132], which does cannot obtain a "question-to-question similarity". However, the modularity characteristic of our architecture will facilitate the use of alternative QA systems, which could improve the precision in the search for the correct answer to a specific question.

Next, we show some examples of information searches extracted from the case study. These examples illustrate the benefits of the services described in our proposal.

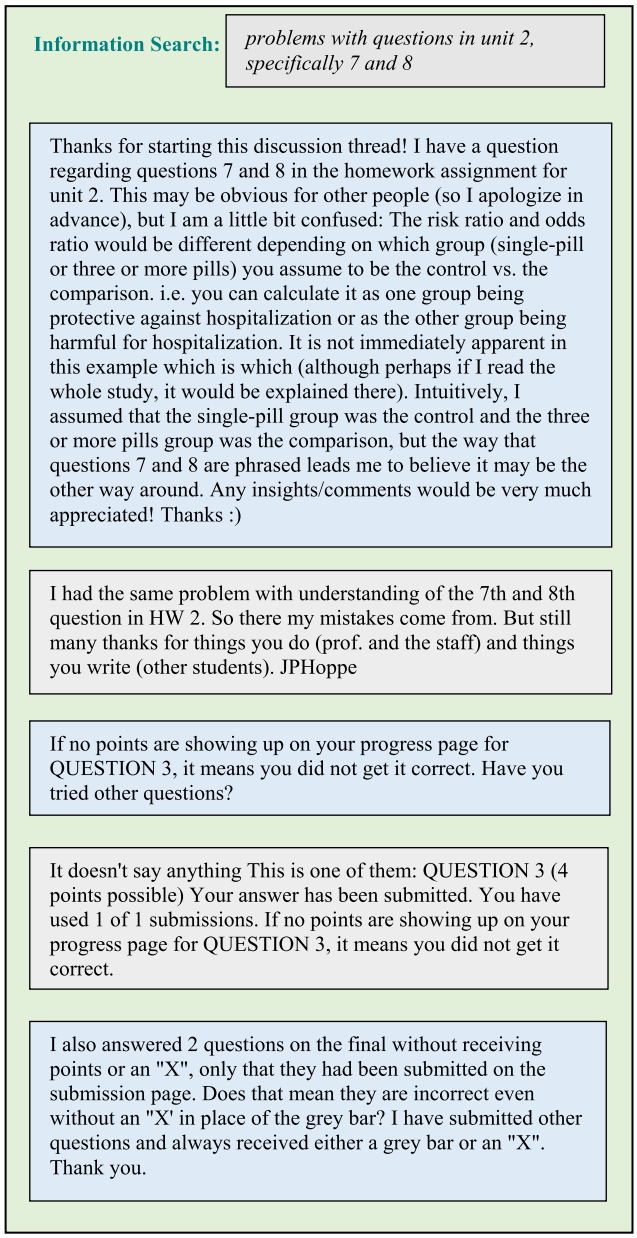

FIGURE 8. An example of an Information Retrieval search.

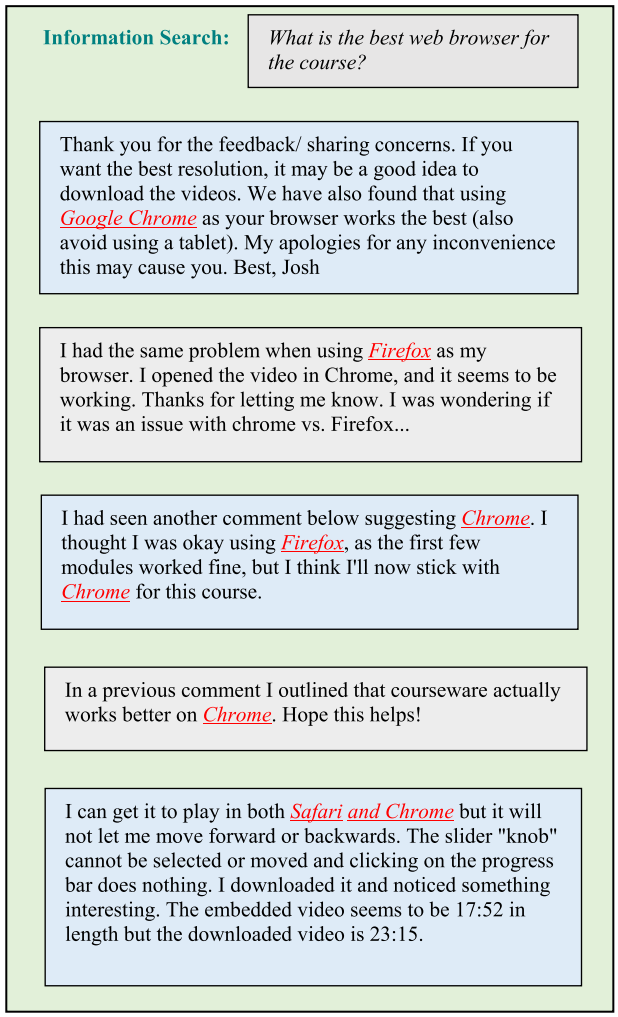

FIGURE 9. An example of a Question Answering search.

The first example shows an instance of a general information search that can be submitted by both teachers and students: ''problems with questions in unit 2, specifically 7 and 8''. In this search, the system extracted the information as an Information Retrieval (IR) engine, that is, it presents both the questions and answers in the forums that are relevant to the information search as shown in Figure 8. In this example, we have run our IR tool presented in [131], where the terms in the questions and forums are enriched with lexical and syntactic knowledge generated by a POS-tagger and a syntactic Chunker. This ensures the inclusion of dependency information between terms that are not considered in traditional IR similarity measures. Specifically, the most outstanding terms extracted from the content of the courses (i.e. ''question'' or ''unit'') are weighted according to the Deviation from Randomness (DFR) measure [133] and properly enhanced with the lexical and syntactic knowledge, as detailed in [131]. The noun phrases generated in ''question 7 and 8 in unit 2'' are enriched with their semantic variations (e.g. ''question'' and ''homework'' share common inherited hypernyms). The results presented in this figure show the first most relevant post that coincides with most of the information search terms, followed by other posts where the number of coinciding terms decrease until the least relevant post that only matches the term ''question''.

The second example in Figure 9 illustrates the types of questions that also extract specific information searched by teachers or students as in Question Answering applications: ''What is the best web browser for the course?''. In this kind of search, the system benefits from a deeper understanding of both the question and the text forums in order to analyses the information search. The aim is to detect the kind of entities that the user is searching for (e.g. a date or a URL), and exploit the semantic information about possible answers to extract the correct answer. For example, in the query

example, our QA tool that has participated in CLEF and TREC competitions, named AliQAn [132], concludes that the user is searching for ''web browsers'' after a pattern matching based on lexical, syntactic and semantic knowledge (''what'' + headOfTheFollowingNounPhrase). In this figure, different web browsers (e.g. ''Google Chrome'') are automatically identified and highlighted in the posts, which are detected in AliQAn by means of semantic knowledge extracted from WordNet and Wikipedia as detailed in Ferrández *et al.* [134]. Next, the post that contains an entity of the required type are sorted by relevance according to the information search.

E. KEY PERFORMANCE INDICATORS (KPI) INFERENCE

This functionality will report different KPIs of the online course extracted from the forums, especially useful for teachers and administrators of the ODL platform. The results will be presented through dashboards. Indicators that express participation, motivation, achievement, and the degree of student drop-out are very useful for managing online courses and designing learning strategies.

The method presented in the work of Peral *et al.* [135] for the identification of relevant KPIs, which has been evaluated in an educational context, could be used for this module. This method consists of 6 stages in which, basically, the business strategy model, the KPIs and the multidimensional model are defined. The method uses DM techniques to identify the KPIs that are really relevant.

In this work, the authors identify indicators such as increment in the number of students, dropout ratio, student recovery ratio, % of active students, % of students who fail the course, % of students passing the exams without seeing the corresponding lessons, and % of students taking the course in a continuous or sequential pattern from a MOOC course. After that, they evaluate over Open Data extracted from the University of Alicante the objective about ''student achievement'' which is measured by the student results. After applying their method and DM techniques, they extract the most determinant indicators and check their relevance.

It is necessary to emphasize other indicators that are extracted automatically from the specific data of the forums and that have not been taken into account previously. In this set, the mentioned method of Mora *et al.* [20] is included. This method is able to extract the main topics discussed in the forums and displays them in a clean, orderly, and classified way with the aim to enrich the student-teacher interaction taking into account the contribution of knowledge among all the information included in the forums.

To conclude our case study, we will show the KPI inference process. KPIs are used to evaluate the success of specific activities, very often connected to companies. In our case, we will use KPIs to identify the students' needs through education. It is worthwhile to note that KPIs will be extracted only from the forums. In fact, it would be interesting to discover other KPIs, such as student progression along the course, percentage of dropouts, etc. The next experiment in a cloud representation will give us the opportunity to understand how

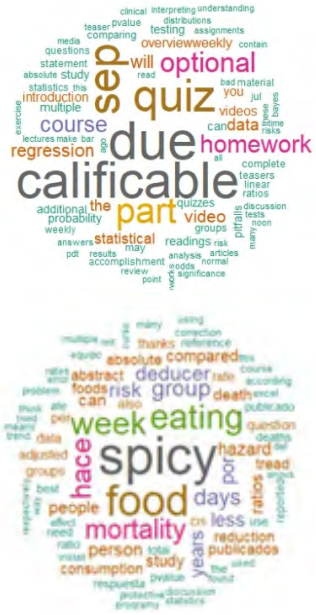

FIGURE 10. Word clouds of ''Course Material Feedback'' and ''External Resources'' forums.

it is possible to obtain these types of indicators and, therefore, will open new directions with the possibilities of KPIs in the education context. In this regard, it will be possible to record feedback from students, teachers, and instructors, to facilitate continuous improvement of each course.

In order to obtain a graphical visualization with the key words of the different forum categories the word cloud (or tag cloud) has been used. In Figure 10 the word clouds of the ''Course Material Feedback'' and ''External Resources'' forums are shown.

As observed in the figure, the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence. For instance, in the ''Course Material Feedback'' forum we can appreciate some key terms: ''due'' (regarding due dates), ''homework'' or ''quiz''. In the ''External Resources'' forum relevant words related to ''food'', ''spicy'' and ''mortality'' can be observed. This is because these words are repeated in several posts about experiments relating to the association of mortality with the spicy food.

Once we have graphically obtained these relevant data, we have defined a three-step process that allows us to confirm the relevant topics previously extracted (using the clustering techniques mentioned in Section III-C) and the finding of new indicators. The process consists of the following steps: (i) identification of the most relevant words in the cloud, (ii) automatic extraction of topics, and (iii) combination of the results of the previous two steps. After the combination of the results, the relevant topics will be confirmed, and new indicators will be discovered. These will enable new patterns of the students' behavior to be obtained, which will help to make decisions that will improve the course.

```
{
 "sentimentAttributes": [
  [
   {
    "subject": "OVERALL",
    "intensity": 0.6666667,
    "emotionLabels": [
     "happy",
     "neutral"
    ],
    "sentimentCategory": "positive"
   },
   {
    "subject": "Francesco",
    "intensity": 0.6666667,
    "emotionLabels": [
     "happy",
     "neutral"
    ],
    "sentimentCategory": "positive"
   }
  ]
 ]
}

{  "text" : ["Thank you! The course was great! Thank you very
much for your effort! I will recommend it. Francesco."],
 "subjects" : [],
 "discoverSubjects" : true }
```


FIGURE 11. Example of post processing with Sentiment Analysis tool.

Applying this process to our case study: (i) With the word cloud of ''Course Material Feedback'' forum we see that the terms ''due'' and ''homework'' are very relevant for students. (ii) After the automatic extraction of topics, if we analyses Table 3 (main topics), three topics related to these terms have been extracted: (1) Homework questions (46.1% of the threads); (2) homework submission problems (6.7%); (3) homework due date? (2.2%). (iii) Combining the data of the cloud with the automatically extracted topics we can affirm that the ''due date'' is a very relevant term due to its appearance frequency in all the posts (according to the cloud) although only 2.2% of the posts have been extracted related to this topic. Subsequently, we can conclude that course managers should take new measures concerning this topic by, for example, announcing in advance and/or periodically the work deadlines dates and publish them via different channels.

To conclude this section, we present the extraction of a KPI about the student satisfaction percentage. This would be an indicator that the objectives of the course are being fulfilled. The course threads about ''Congratulations'' topic (11.8% of the total threads) have been processed with a Sentiment Analysis and Opinion Mining tool in order to discover the polarity of the opinions expressed in these threads (positive, negative or neutral). These systems determine whether a message (or a fragment of it) expresses a positive, negative, or neutral sentiment.

We have used the GPLSI system: supervised sentiment analysis in Twitter [136], [137], submitted for the SemEval 2014 Task 9 (Sentiment Analysis in Twitter). It consists of a supervised approach using machine learning techniques, without employing any external knowledge and resources. It uses the term in the dataset as features. These terms are combined to create skipgrams (not-adjacent ngrams) that are employed as features for a supervised machine learning algorithm. In Figure 11, the processing of thread #155 of ''Course Material Feedback'' forum with the GPLSI system is shown.

In Figure 11, the post text (''Thank you! The course was great! Thank you very much for your effort! I will recommend it. Francesco.'') can be observed. The polarity (sentimentCategory: positive) and intensity (0. 6666667) of all text (tagged as OVERALL) are identified. Furthermore, the polarity and intensity about automatically extracted subjects (Francesco) are presented.

Of all the threads related to ''Congratulations'' topic, 95.25% had a positive sentiment and 4.75% had a neutral one. This indicator is very important, showing that the student satisfaction percentage is very high and the general opinion of the students about the course is very positive.

IV. CONCLUSIONS AND FUTURE WORK

The ODL platforms have become very popular in recent years. Forums are a central communication tool in many courses included in online educational platforms. These courses rely mainly on discussion forums for interaction among students. However, the learning advantages that these tools should provide are very often not exploited. Forums do not support learning if many messages are produced, especially when they are posted in a disordered and unstructured way which makes it difficult and time consuming for the user to analyses the information.

Numerous studies have been performed to look for information about how students search for and manage information as well as other aspects of forum operations to improve management and learning effects. In this paper we have reviewed the approaches and techniques related to online courses management to discover what challenges need to be resolved. We conclude that no effective solutions have been proposed and there are three main unresolved challenges: (1) the efficient management and monitoring of massive forums; (2) the effective search mechanisms over questions and answers present in the forums; (3) the extraction of relevant Key Performance Indicators for improving the learning and teaching processes.

Furthermore, we have presented a set of analytics techniques and a general architecture with three basic functionalities that resolve the abovementioned challenges: (1) the use of automatic tools and applications to facilitate the tracking and monitoring of online learning communities by using NLP techniques (lexical, syntactic and semantic analysis of the text, anaphora and ellipsis resolution, etc.) and clustering; (2) the application of Information Retrieval and Question Answering techniques for searching information; (3) the employment of Data Mining techniques to extract the relevant KPIs.

The architecture defined in this work provides the main advantage of a system that has the ability to solve or tackle the main challenges in forums. On the one hand, with the topic detection, the clustering and the use of Data Mining, the irrelevant material can be eliminated. On the other hand, the NLP techniques (dictionaries, parsing and linguistic phenomena resolution) allows the system to work with and correct text written in an informal language with grammatical and typographic errors. The benefits of our proposal have been shown on an online course of Stanford University named Statistics in Medicine (MEDSTATS).

Future research lines will involve an assessment of the impact of this study through a deeper evaluation of the architecture, with a comparative analysis of the interaction between alternative tools (e.g. different QA systems), and on a set of different and larger data sets. The intention is to strengthen the use of ODL platforms forums, allowing them to boost student learning, as well as serving as a tool for teachers to monitor the learning process of their students. The aim is to overcome the current drawbacks of ODL and they become the basis of new teaching-learning processes such as e-learning, collaborative learning and project-based learning. In this way, we will be able to approach the ideal goal of personalized education, in which each student advances according to his/her characteristics and interests.

APPENDIX

MEDSTATS Course Program. Unit 1: Descriptive Statistics and Looking at Data

Module 1: Introduction to datasets
Module 2: Types of data
Module 3: Visualizing data
Module 4: Measures of central tendency (mean, median)
Module 5: Dispersion of the data (standard deviation, percentiles)
Module 6: Exploring real data: lead in lipstick
Unit 1 Homework
Additional Readings (optional)
Unit 1 R exercise 1 (Optional)

Unit 2: Review of Study Designs; Measures of Disease Risk and Association

Module 1: Review of study designs
Module 2: Measures of disease frequency
Module 3: Absolute risk differences
Module 4: Relative risks (rate ratios, risk ratios, hazard ratios, odds ratios)
Module 5: Odds ratios can mislead
Module 6: Communicating risks clearly: absolute vs. relative risks
Unit 2 Homework
Additional Readings (optional)

Unit 3: Probability, Bayes' Rule, Diagnostic Testing

Module 1: Basic Probability
Module 1 Optional: Calculating Probabilities: Permutations and Combinations
Module 2: Rules of Probability
Module 3: Probability Trees and Conditional Probability
Module 4: Bayes' Rule
Module 4 Optional: Conditional probability, Bayes' rule, and the odds ratio
Module 5: Diagnostic testing
Unit 3 Homework

Unit 4: Probability Distributions

Module 1: Probability Distributions
Module 2: Expected Value
Module 3: Variance
Module 4: The Binomial Distribution
Module 5: The Normal and Standard Normal Distributions
Module 6: The Normal Approximation to the Binomial
Module 7: Assessing Normality in Data
Unit 4 Homework

Unit 5: Statistical Inference

Module 1: Review of Z-distribution, Introduction to T-distribution
Module 2: Introduction to Statistical Inference
Module 3: Introduction to the Distribution of a Statistic
Module 4: Distributions of some common Statistics
Module 5: Confidence Intervals (estimation)
Module 6: Where does the Margin of Error come from in Polls?
Module 7: Hypothesis Testing (p-values)
Module 8: HIV Vaccine Trial/Bayesian Inference
Unit 5 Homework
Additional Readings (optional)

Unit 6: P-values (errors, statistical power, and pitfalls)

Module 1: Type I and Type II errors and Statistical power
Module 1 Optional: Sample Size Formulas, Derivations
Module 2: P-value pitfalls: Statistical vs. Clinical Significance
Module 3: P-value pitfalls: Multiple Testing
Guest Lecture - Case Study: Multiple Testing in Cardiovascular Medicine (optional)
Module 4: P-value pitfalls: Don't compare P-values!
Module 5: P-value pitfalls: Failure to prove an effect is not proof of no effect
Module 6: P-value pitfalls: Correlation is not Causation
Module 7: Introduction to Correlated Data
Module 8: Overview of Statistical Tests
Unit 6 Homework
Additional Readings (optional)

Unit 7: Statistical Tests

Module 1: Comparing Means between 2 Groups
Module 2: Comparing Means between more than 2 Groups
Module 3: Alternative tests to the ttest and ANOVA (non-parametric tests)
Module 4: Comparing Proportions between 2 Groups
Module 5: Comparing Proportions between more than 2 Groups
Module 6: Comparing Time-to-Event Outcomes between 2 or more Groups
Unit 7 Homework
Additional Readings (optional)

Unit 7 R exercise (Optional)

Unit 8: Regression Analysis

Module 1: Covariance and Correlation

Module 2: Simple Linear Regression

Module 3: Residual Analysis

Module 4: Multiple Linear Regression and Statistical Adjustment

Module 5: Categorical Predictors in Regression

Module 6: Practice Interpreting Linear Regression Results

Module 7: Regression Worries: Overfitting and Missing Data

Module 7 Optional: Variable Transformation

Unit 8 Homework

Additional Readings (optional)

Unit 8 R exercise (Optional)

Unit 9: Regression II: Logistic Regression, Cox Regression

Module 1: Logistic Regression

Module 2: Practical Example: Interpreting results from Logistical Regression

Module 3: Testing the ''linear in the logit'' Assumption of Logistical Regression

Module 4: Interactions

Module 5: Introduction to Cox Regression

Module 6: Regression Worries: Residual Confounding

Unit 9 Homework

REFERENCES

[1] *Open and Distance Learning: Trends, Policy and Strategy Considerations*, UNESCO, Paris, France, 2002.

[2] H. Mora-Mora, M. T. Signes-Pont, and G. De Miguel Casado, ''Information search habits of first year college students,'' *Int. J. Knowl. Soc. Res.*, vol. 5, no. 4, pp. 26–34, 2014.

[3] M. Masud, ''Knowledge update in collaborative knowledge sharing systems,'' *Int. J. Knowl. Soc. Res.*, vol. 6, no. 3, pp. 19–31, 2015.

[4] V. L. L. De Azevedo and M. Borges, ''More collaboration, more collective intelligence,'' *Int. J. Knowl. Soc. Res.*, vol. 6, no. 3, pp. 1–18, 2015.

[5] M. D. Lytras, H. I. Mathkour, H. Abdalla, W. Al-Halabi, C. Yanez-Marquez, and S. W. M. Siqueira, ''An emerging–social and emerging computing enabled philosophical paradigm for collaborative learning systems: Toward high effective next generation learning systems for the knowledge society,'' *Comput. Hum. Behav.*, vol. 51, pp. 557–561, Oct. 2015.

[6] C. B. Mahmoud, I. Azaiez, F. Bettahar, and F. Gargouri, ''Discovery mechanism for learning semantic Web service,'' in *Web Services: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2019, pp. 575–596.

[7] H. M. Mora, M. T. S. Pont, G. De Miguel Casado, and V. G. Iglesias, ''Management of social networks in the educational process,'' *Comput. Hum. Behav.*, vol. 51, pp. 890–895, Oct. 2015.

[8] D. F. Onah, J. R. Sinclair, R. Boyatt, and J. Foss, ''Massive open online courses: Learner participation,'' in *Proc. 7th Int. Conf. Educ., Res. Innov. (iCERi)*, 2014, pp. 2348–2356.

[9] H. Lentell and J. O'Rourke. ''Tutoring large numbers: An unmet challenge,'' *Int. Rev. Res. Open Distance Learn.*, vol. 5, no. 1, pp. 1–17, 2004.

[10] D. Hoogeveen, L. Wang, T. Baldwin, and K. M. Verspoor, ''Web forum retrieval and text analytics: A survey,'' *Found. Trends Inf. Retr.*, vol. 12, no. 1, pp. 1–163, 2018.

[11] Al-A. Abri, Y. Jamoussi, N. Kraiem, and Z. Al-Khanjari, ''Comprehensive classification of collaboration approaches in E-learning,'' *Telematics Inform.*, vol. 34, no. 6, pp. 878–893, 2017.

[12] A. Ramesh, D. Goldwasser, B. Huang, H. Daume, and L. Getoor, ''Understanding MOOC discussion forums using seeded LDA,'' in *Proc. ACL Workshop Innov. Use NLP Building Educ. Appl.*, 2014, pp. 28–33.

[13] A. I. Obasa, N. Salim, and A. Khan, ''Hybridization of bag-of-words and forum metadata for Web forum question post detection,'' *Indian J. Sci. Technol.*, vol. 8, no. 32, pp. 1–12, 2016.

[14] L. Hong and B. D. Davison, ''A classification-based approach to question answering in discussion boards,'' in *Proc. 32nd Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2009, pp. 171–178.

[15] R. Catherine, A. Singh, R. Gangadharaiah, D. Raghu, and K. Visweswariah, ''Does similarity matter? The case of answer extraction from technical discussion forums,'' in *Proc. 24th Int. Conf. Comput. Linguistics (COLING)*, 2012, pp. 175–184.

[16] R. Catherine, R. Gangadharaiah, K. Visweswariah, and D. Raghu, ''Semi-supervised answer extraction from discussion forums,'' in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 1–9.

[17] J. Huang, M. Zhou, and D. Yang, ''Extracting Chatbot knowledge from online discussion forums,'' in *Proc. 20th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2007, pp. 423–428.

[18] P. Martínez-Barco *et al.*, ''LEGOLANG: Técnicas de deconstrucción aplicadas a las tecnologías del Lenguaje Humano,'' *Procesamiento Lenguaje Natural*, vol. 51, pp. 219–222, Sep. 2013.

[19] A. Blum, and T. Mitchell, ''Combining labeled and unlabeled data with co-training,'' in *Proc. 11th Annu. Conf. Comput. Learn. Theory (COLT)*, 1998, pp. 92–100.

[20] D. Feng, E. Shaw, J. Kim, and E. Hovy, ''Learning to detect conversation focus of threaded discussions,'' in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2006, pp. 208–215.

[21] H. Mora, A. Ferrández, D. Gil, and J. Peral, ''A computational method for enabling teaching-learning process in huge online courses and communities,'' *Int. Rev. Res. Open Distrib. Learn.*, vol. 18, no. 1, pp. 225–246, 2017.

[22] P. Shea and T. Bidjerano, ''Community of inquiry as a theoretical framework to foster 'epistemic engagement' and 'cognitive presence' in online education,'' *Comput. Educ.*, vol. 52, no. 3, pp. 543–553, 2009.

[23] M. De Laat, V. Lally, L. Lipponen, and R. J. Simon, ''Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis,'' *J. Comput. Supported Collaborative Learn.*, vol. 2, no. 1, pp. 87–103, 2007.

[24] H.-J. Suh and S.-W. Lee, ''Collaborative learning agent for promoting group interaction,'' *ETRI J.*, vol. 28, no. 4, pp. 461–474, 2006.

[25] K. Swan, P. Shea, E. Fredericksen, A. Pickett, W. Pelz, and G. Maher, ''Building knowledge building communities: Consistency, contact and communication in the virtual classroom,'' *J. Educ. Comput. Res.*, vol. 23, no. 4, pp. 359–383, 2000.

[26] J. A. Baxter and J. Haycock, ''Roles and student identities in online large course forums: Implications for practice,'' *Int. Rev. Res. Open Distrib. Learn.*, vol. 15, no. 1, 2014.

[27] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé, ''Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses,'' in *Proc. NIPS Data-Driven Educ. Workshop*, vol. 11, 2013, p. 14.

[28] C. Romero, M. I. López, J. M. Luna, and S. Ventura, ''Predicting students' final performance from participation in on-line discussion forums,'' *Comput. Educ.*, vol. 68, pp. 458–472, Oct. 2013.

[29] E. Z.-F. Liu, S.-S. Cheng, and C. H. Lin, ''The effects of using online Q&A discussion forums with different characteristics as a learning resource,'' *Asia-Pacific Edu. Res.*, vol. 22, no. 4, pp. 667–675, 2013.

[30] B. De Wever, T. Schellens, M. Valcke, and H. Van Keer, ''Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review,'' *Comput. Educ.*, vol. 46, no. 1, pp. 6–28, 2006.

[31] J. Chávez, R. Montaño, and R. Barrera, ''Structure and content of messages in an online environment: An approach from participation,'' *Comput. Hum. Behav.*, vol. 54, pp. 560–568, Jan. 2016.

[32] J. Brace-Govan, ''A method to track discussion forum activity: The moderators' assessment matrix,'' *Internet Higher Educ.*, vol. 6, no. 4, pp. 303–325, 2003.

[33] S. Premagowrie, R. V. Kalai, and R. C. Ho, ''Online forum: A platform that affects students' learning,'' *Amer. Int. J. Social Sci.*, vol. 3, no. 7, pp. 107–116, 2014.

[34] V. P. Dennen, ''Looking for evidence of learning: Assessment and analysis methods for online discourse,'' *Comput. Hum. Behav.*, vol. 24, no. 2, pp. 205–219, 2008.

[35] W. McKenzie and D. Murphy, '''I hope this goes somewhere': Evaluation of an online discussion group,'' *Australas. J. Educ. Technol.*, vol. 16, no. 3, pp. 239–257, 2000.

[36] M. Guzdial and J. Turns, "Effective discussion through a computer-mediated anchored forum," *J. Learn. Sci.*, vol. 9, no. 4, pp. 437–469, 2000.
[37] G. S. Stump, J. DeBoer, J. Whittinghill, and L. Breslow, "Development of a framework to classify MOOC discussion forum posts: Methodology and challenges," in *Proc. NIPS Workshop Data Driven Educ.*, 2013, pp. 1–20.
[38] C. K. Coursaris and M. Liu, "An analysis of social support exchanges in online HIV/AIDS self-help groups," *Comput. Hum. Behav.*, vol. 25, no. 4, pp. 911–918, 2009.
[39] S. W. Thomas, "Mining software repositories with topic models," School Comput., Queen's Univ., Kingston, ON, Canada, Tech. Rep. 2012-586, 2012.
[40] L. Fan, Y. Zhang, Y. Dang, and H. Chen, "Analyzing sentiments in Web 2.0 social media data in Chinese: experiments on business and marketing related Chinese Web forums," *Inf. Technol. Manage.*, vol. 14, no. 3, pp. 231–242, 2013.
[41] R. Anbalagan, A. Kumar, and K. Bijlani, "Footprint model for discussion forumsin MOOC," *Procedia Comput.Sci*., vol. 58, pp. 530–537, Jan. 2015.
[42] N. Yusof and A. A. Rahman, "Students' interactions in online asynchronous discussion forum: A social network analysis," in *Proc. Int. Conf. Educ. Technol. Comput. (ICETC)*, vol. 9, 2009, pp. 25–29.
[43] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decis. Support Syst.*, vol. 48, no. 2, pp. 354–368, 2010.
[44] S. D'Mello, A. Olney, and N. Person. "Mining collaborative patterns in tutorial dialogues," *J. Educ. Data Mining*, vol. 2, no. 1, pp. 1–37, 2010.
[45] W. L. Cade, J. L. Copeland, N. K. Person, and S. K. D'Mello, "Dialogue modes in expert tutoring," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 470–479.
[46] R. S. Hoover and A. L. Koerber, "Using NVivo to answer the challenges of qualitative research in professional communication: Benefits and best practices tutorial," *IEEE Trans. Prof. Commun.*, vol. 54, no. 1, pp. 68–82, Mar. 2011.
[47] G. Wang, "Research on hotspot discovery in Internet public opinions based on improved K-means," *Comput. Intell. Neurosci.*, vol. 2013, p. 5, Jan. 2013.
[48] A. Hazem, B. E. A. Boussaha, and N. Hernández, "MappSent: A textual mapping approach for question-to-question similarity," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2017, pp. 291–300.
[49] X. Li and D. Roth, "Learning question classifiers," in *Proc. 19th Int. Conf. Comput. Linguistics*, vol. 1, 2002, pp. 1–7.
[50] Y. Liu, S. Li, Y. Cao, C.-Y. Lin, D. Han, and Y. Yu, "Understanding and summarizing answers in community-based question answering services," in *Proc. 22nd Int. Conf. Comput. Linguistics (COLING)*, 2008, pp. 497–504.
[51] D. Metzler and W. B. Croft,"Analysis of statistical question classification for fact-based questions," *Inf. Retr.*, vol. 8, no. 3, pp. 481–504, 2005.
[52] J. Suzuki, H. Taira, Y. Sasaki, and E. Maeda, "Question classification using HDAG kernel," in *Proc. ACL Workshop Multilingual Summarization Question Answering*, vol. 12, 2003, pp. 61–68.
[53] Y. Li, L. Su, J. Chen, and L. Yuan, "Semi-supervised learning for question classification in CQA," *Natural Comput.*, vol. 16, no. 4, pp. 567–577, 2016.
[54] C. Pechsiri and R. Piriyakul, "developing a why–how question answering system on community Web boards with a causality graph including procedural knowledge," *Inf. Process. Agricult.*, vol. 3, no. 1, pp. 36–53, 2016.
[55] J. He and D. Dai, "Summarization of yes/no questions using a feature function model," in *Proc. 3rd Asian Conf. Mach. Learn. (ACML)*, 2011, pp. 351–366.
[56] S. M. Harabagiu *et al.*, "FALCON: Boosting knowledge for answer engines," in *Proc. 9th Text Retr. Conf. (TREC)*, 2000, pp. 479–488.
[57] E. Hovy, L. Gerber, U. Hermjakob, C.-Y. Lin, and D. Ravichandran, "Toward semantics-based answer pinpointing," in *Proc. Meeting North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2001, pp. 1–7.
[58] A. Ittycheriah, M. Franz, W.-J. Zhu, A. Ratnaparkhi, and R. J. Mammone, "Question answering using maximum entropy components," in *Proc. 2nd Meeting North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2001, pp. 1–7.
[59] E. M. Voorhees, "The TREC-8 question answering track report," in *Proc. 8th Text Retr. Conf. (TREC)*, 1999, pp. 77–82.
[60] S. Lytinen and N. Tomuro, "The use of question types to match questions in FAQ finder," in *Proc. AAAI Spring Symp. Mining Answers Texts Knowl. Bases (SS)*, 2002, pp. 46–53.
[61] J. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artif. Intell. Rev.*, vol. 35, no. 2, pp. 137–154, 2011.
[62] Z. Huang, M. Thint, and Z. Qin, "Question classification using head words and their hypernyms," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2008, pp. 927–936.
[63] L. Chen, D. Zhang, and L. Mark, "Understanding user intent in community question answering," in *Proc. 21st Int. World Wide Web Conf.*, 2012, pp. 823–828.
[64] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein, "Exploring question subjectivity prediction in community QA," in *Proc. 31st Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 735–736.
[65] N. Aikawa, T. Sakai, and H. Yamana, "Community QA question classification: Is the asker looking for subjective answers or not?" *IPSJ Online Trans.*, vol. 4, no. 2, pp. 1–9, Jul. 2011.
[66] F. M. Harper, D. Moy, and J. A. Konstan, "Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2009, pp. 759–768.
[67] H. Amiri, Z.-J. Zha, and T.-S. Chua, "A pattern matching based model for implicit opinion question identification," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 46–52.
[68] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
[69] H. Fu *et al.*, "ASELM: Adaptive semi-supervised ELM with application in question subjectivity identification," Neurocomputing, vol. 207, pp. 599–609, Sep. 2016.
[70] B. Li, Y. Liu, and E. Agichtein, "CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2008, pp. 937–946.
[71] I. Gurevych, D. Bernhard, K. Ignatova, and C. Toprak, "Educational question answering based on social media content," in *Proc. Int. Conf. Artif. Intell. Educ. (IJAIED)*, 2009, pp. 133–140.
[72] P. Biyani, "Analyzing subjectivity and sentiment of online forums," Ph.D. dissertation, Inf. Sci. Technol., Pennsylvania State Univ., State College, PA, USA, 2014.
[73] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. 1st ACM Int. Conf. Web Search Data Mining (WSDM)*, 2008, pp. 183–194.
[74] L. T. Le, C. Shah, and E. Choi, "Evaluating the quality of educational answers in community question-answering," in *Proc. 16th ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL)*, 2016, pp. 129–138.
[75] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proc. 31st Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 483–490.
[76] M. Lui and T. Baldwin, "You are what you post: User-level features in threaded discourse," in *Proc. 14th Australas. Document Comput. Symp. (ADCS)*, 2009, pp. 98–105.
[77] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar, "Automatic scoring of online discussion posts," in *Proc. 2nd ACM Workshop Inf. Credibility Web (WICOW)*, 2008, pp. 19–26.
[78] M. Weimer, I. Gurevych, and M. Mühlhäuser, "Automatically assessing the post quality in online discussions on software," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics Interact. Poster Demonstration Sessions*, 2007, pp. 125–128.
[79] M. Weimer and I. Gurevych, "Predicting the perceived quality of Web forum posts," in *Proc. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2007, pp. 643–648.
[80] G. Burel, Y. He, and H. Alani, "Automatic identification of best answers in online enquiry communities," in *Proc. Extended Semantic Web Conf. (ESWC)*, 2012, pp. 514–529.
[81] G. Burel, "Community and thread methods for identifying best answers in online question answering communities," Ph.D. dissertation, Knowl. Media Inst., Open Univ., Milton Keynes, U.K., 2016.
[82] A. Y. K. Chua and S. Banerjee, "Measuring the effectiveness of answers in Yahoo! Answers," *Online Inf. Rev.*, vol. 39, no. 1, pp. 104–118, 2015.
[83] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Exploiting user feedback to learn to rank answers in Q&A forums: A case study with stack overflow," in *Proc. 36th Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2013, pp. 543–552.

[84] G. Dror, Y. Maarek, and I. Szpektor, "Will my question be answered? Predicting 'question answerability' in community question-answering sites," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, 2013, pp. 499–514.

[85] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the past: Answering new questions with past answers," in *Proc. 21st Int. World Wide Web Conf.*, 2012, pp. 759–768.

[86] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "Finding the right facts in the crowd: Factoid question answering over social media," in *Proc. 17th Int. World Wide Web Conf.*, 2008, pp. 467–476.

[87] P. Nakov *et al.*, "SemEval-2017 task 3: Community question answering," in *Proc. 11th Int. Workshop Semantic Eval. Assoc. Comput. Linguistics (SemEval)*, 2017, pp. 1–5.

[88] B. Dom, and D. Paranjpe, "A Bayesian technique for estimating the credibility of question answerers," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2008, pp. 399–409.

[89] B. V. Hanrahan, G. Convertino, and L. Nelson, "Modeling problem difficulty and expertise in stackoverflow," in *Proc. ACM Conf. Companion Comput. Supported Cooperat. Work Social Comput. (CSCW)*, 2012, pp. 91–94.

[90] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums: the case of Yahoo! Answers," in *Proc. 14th ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 866–874.

[91] L. Guo and X. Hu, "Identifying authoritative and reliable contents in community question answering with domain knowledge," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, 2013, pp. 133–142.

[92] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proc. 16th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2007, pp. 919–922.

[93] P. Jurczyk and E. Agichtein, "Hits on question answer portals: Exploration of link analysis for author ranking," in *Proc. 30th Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2007, pp. 845–846.

[94] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang, "Quality-aware collaborative question answering: methods and evaluation," in *Proc. 2nd ACM Int. Conf. Web Search Data Mining (WSDM)*, 2009, pp. 142–151.

[95] G. Zhou, S. Lai, K. Liu, and J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 1662–1666.

[96] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "ExpertRank: A topic-aware expert finding algorithm for online knowledge communities," *Decis. Support Syst.*, vol. 54, no. 3, pp. 1442–1451, 2013.

[97] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proc. 16th Int. World Wide Web Conf.*, 2007, pp. 221–230.

[98] Z. Zhao, L. Zhang, X. He, and W. Ng, "Expert finding for question answering via graph regularized matrix completion," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 993–1004, Apr. 2015.

[99] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha, "Learning to recognize reliable users and content in social media with coupled mutual reinforcement," in *Proc. 18th Int. World Wide Web Conf.*, 2009, pp. 51–60.

[100] S. Xie *et al.*, "Effective crowd expertise modeling via cross domain sparsity and uncertainty reduction," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2016, pp. 648–656.

[101] M. Fu, M. Zhu, Y. Su, Q. Zhu, and M. Li, "Modeling temporal behavior to identify potential experts in question answering communities," in *Proc. Int. Conf. Cooperat. Design, Vis. Eng. (CDVE)*, 2016, pp. 51–58.

[102] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: Stackoverflow," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, 2013, pp. 886–893.

[103] A. Pal, S. Chang, and J. A. Konstan, "Evolution of experts in question answering communities," in *Proc. 6th AAAI Int. Conf. Weblogs Social Media (ICWSM)*, 2012, pp. 274–281.

[104] J. Wang, J. Sun, H. Lin, H. Dong, and S. Zhang, "Predicting best answerers for new questions: An approach leveraging convolution neural networks in community question answering," in *Proc. Chin. Nat. Conf. Social Media Process.*, 2016, pp. 29–41.

[105] P. Nakov *et al.*, "SemEval-2016 task 3: Community question answering," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 1–6.

[106] A. Barrón-Cedeno *et al.*, "ConvKN at semeval-2016 task 3: Answer and question selection for question answering on Arabic and English fora," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 896–903.

[107] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. (2014). "A convolutional neural network for modelling sentences." [Online]. Available: https://arxiv.org/abs/1404.2188

[108] S. Filice, D. Croce, A. Moschitti, and R. Basili, "Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 1116–1123.

[109] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[110] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[111] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 3111–3119.

[112] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[113] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, and A. Y. Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 801–809.

[114] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 3294–3302.

[115] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–19.

[116] S. Arora, L. Yingyu, and M. Tengyu, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. 17th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–16.

[117] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, pp. 1–26.

[118] S. Rho, W. Rahayu, and U. T. Nguyen, "Advanced issues on topic detection, tracking, and trend analysis for social multimedia," *Adv. Multimedia*, vol. 2015, pp. 1–2, Mar. 2015.

[119] J. Allan, S. Harding, D. Fisher, A. Bolivar, S. Guzman-Lara, and P. Amstutz, "Taking topic detection from evaluation to practice," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2005, p. 101a.

[120] G. N. Hu, X. Y. Dai, Y. Y. Song, S. Huang, and J. Chen, "A synthetic approach for recommendation: Combining ratings, social relations, and reviews," in *Proc. Int. Conf. Artif. Intell. (IJCAI)*, 2015, pp. 1756–1762.

[121] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[122] P. Winoto, T. Y. Tang, and G. I. McCalla, "Contexts in a paper recommendation system with collaborative filtering," *Int. Rev. Res. Open Distrib. Learn.*, vol. 13, no. 5, pp. 56–75, 2012.

[123] J. Sathick and J. Venkat, "A generic framework for extraction of knowledge from social Web sources (social networking websites) for an online recommendation system," *Int. Rev. Res. Open Distrib. Learn.*, vol. 16, no. 2, pp. 247–271, 2015.

[124] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 497–506.

[125] Y. Feng, H. Fani, E. Bagheri, and J. Jovanovic, "Lexical semantic relatedness for Twitter analytics," in *Proc. IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2015, pp. 202–209, doi: 10.1109/ICTAI.2015.41.

[126] F. Kalloubi, E. H. Nfaoui, and O. El Beqqali, "Harnessing semantic features for large-scale content-based hashtag recommendations on microblogging platforms," *Int. J. Semantic Web Inf. Syst.* vol. 13, no. 1, pp. 63–81, 2017.

[127] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2012, pp. 1–5.

[128] A. Ferrández, M. Palomar, and L. Moreno, "An empirical approach to Spanish anaphora resolution," *Mach. Transl.*, vol. 14, nos. 3–4, pp. 191–216, 1999.

[129] M. Palomar *et al.*, "An algorithm for anaphora resolution in Spanish texts," *Comput. Linguistics*, vol. 27, no. 4, pp. 545–567, 2001.

[130] A. F. Wise, Y. Cui, W. Jin, and J. Vytasek, "Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling," *Internet Higher Educ.*, vol. 32, pp. 11–28, Jan. 2017.

[131] A. Ferrández, "Lexical and syntactic knowledge for information retrieval," *Inf. Process. Manage.*, vol. 47, no. 5, pp. 692–705, 2011.
[132] R. Muñoz-Terol *et al.*, "Integrating logic forms and anaphora resolution in the aliqan system," in *Evaluating Systems for Multilingual and Multimodal Information Access* (Lecture Notes in Computer Science), vol. 5706. Berlin, Germany: Springer, 2009, pp. 438–441.
[133] G. Amati, C. Carpineto, and C. Romano, "Comparing weighting models for monolingual information retrieval," in *Comparative Evaluation of Multilingual Information Access Systems* (Lecture Notes in Computer Science), vol. 3237. Berlin, Germany: Springer, 2004, pp. 310–318.
[134] S. Ferrández, A. Toral, O. Ferrández, A. Ferrández and R. Muñoz, "Exploiting wikipedia and eurowordnet to solve cross-lingual question answering," *Inf. Sci.*, vol. 179, no. 20, pp. 3473–3488, 2009.
[135] J. Peral, A. Maté, and M. Marco, "Application of data mining techniques to identify relevant key performance indicators," *Comput. Standards Interfaces*, vol. 50, pp. 55–64, Nov. 2017.
[136] J. Fernández, Y. Gutiérrez, J. M. Gómez, and P. Martínez-Barco, "GPLSI: Supervised sentiment analysis in Twitter using skipgrams," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 294–299.
[137] Y. Gutierrez, D. Tomás, and J. Fernández, "Benefits of using ranking skip-gram techniques for opinion mining approaches," in *Proc. eChallenges e-Conf.*, Nov. 2015, pp. 1–10.

JESÚS PERAL received the Ph.D. degree in computer science from the University of Alicante, in 2001, where he is currently an Assistant Professor with the Department of Software and Computing Systems. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has published many papers (more than 40 papers) in journals and conferences related to his research interests. His main research topics include natural language processing, information extraction, information retrieval, question answering, data warehouses, and business intelligence applications.

ANTONIO FERRÁNDEZ received the Ph.D. degree in computer science from the University of Alicante, in 1998, where he is currently an Assistant Professor with the Department of Software and Computing Systems. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has participated in many conferences, and most of his work has been published in international journals and conferences, with more than 70 published papers. His research topics include information extraction, information retrieval, question answering, natural language processing, and ellipsis and anaphora resolution.

HIGINIO MORA received the B.S. degree in computer science engineering, the B.S. degree in business studies, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 1996, 1997, and 2003, respectively, where he has been a member of the Faculty at the Computer Technology and Computation Department, since 2002, and is currently an Associate Professor and a Researcher of the Specialized Processors Architecture Laboratory. He has participated in many conferences, and most of his work has been published in international journals and conferences, with more than 60 published papers. His research interests include computer modeling, computer architectures, high-performance computing, embedded systems, the Internet of Things, and cloud computing paradigm.

DAVID GIL is currently an Assistant Professor with the Department of Computing Technology and Data Processing, University of Alicante. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has participated in many conferences, and most of his work has been published in international journals and conferences, with more than 50 published papers. His main research topics include artificial intelligence applications, data mining, open data, big data, and decision support system in medical and cognitive sciences.

ERICK KAUFFMANN received the B.S. degree in computer science engineering and the master's degree in computer science from the Technological Institute of Costa Rica, in 1993 and 1999, respectively. He is currently pursuing the Ph.D. degree in computer science with the University of Alicante, Spain. He has been a Professor of the course Information Technology with the Industrial Engineering Department, University of Costa Rica, since 2010. He has participated in numerous national and international projects. His main research topics include natural language processing, information retrieval, and genetic algorithms.