IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA MINING METHODS FOR SOCIAL COMPUTING

Received January 30, 2020, accepted February 12, 2020, date of publication February 19, 2020, date of current version February 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974983

Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach

DIEGO BUENAÑO-FERNANDEZ[1], MARIO GONZÁLEZ[1], DAVID GIL[2], AND SERGIO LUJÁN-MORA[3]

[1]Faculty of Engineering and Applied Sciences, Universidad de Las Américas (UDLA), Quito 170504, Ecuador
[2]Department of Computer Technology and Computation, University of Alicante, 03690 Alicante, Spain
[3]Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain

Corresponding author: Diego Buenaño-Fernandez (diego.buenano@udla.edu.ec)

This work was supported in part by the Universidad de Las Américas, Quito, Ecuador, under Project SIS.DBF.19.02 and Project SIS.MGR.20.02, in part by the Spanish Ministry of Science, Innovation and Universities through the Project ECLIPSE-UA under Grant RTI2018-094283-B-C32, and in part by the Lucentia AGI Grant.

ABSTRACT The large amount of text that is generated daily on the web through comments on social networks, blog posts and open-ended question surveys, among others, demonstrates that text data is used frequently, and therefore; its processing becomes a challenge for researchers. The topic modeling is one of the emerging techniques in text mining; it is based on the discovery of latent data and the search for relationships among text documents. In this paper, the objective of the research is to evaluate a generic methodology based on topic modeling and text network modeling, that allows researchers to gather valuable information from surveys that use open-ended questions. To achieve this, this methodology has been evaluated through the use of a case study in which the responses to a teacher self-assessment survey in an Ecuadorian university have been studied. The main contribution of the article is the inclusion of clustering algorithms in order to complement the results obtained when executing topic modeling. The proposed methodology is based on four phases: (a) Construction of a text database, (b) Text mining and topic modeling, (c) Topic network modeling and (d) The relevance of the identified topics. In previous works, it has been observed that the human interpretative contribution plays an important role in the process, especially in phases (a) and (d). For this reason, the visualization interfaces, such as graphs and dendograms, are of critical importance for researchers in order allow topic to efficiently analyze the results of the topic modeling. As a result of this case study, a compendium of the main strategies that teachers carry out in their classes with the aim of improving student retention is presented. In addition, the proposed methodology can be extended to the analysis of the unstructured textual information found in blogs, social networks, forums, etc.

INDEX TERMS Latent Dirichlet allocation, open-ended questions, teacher self-assessment, topic modeling, topic network.

I. INTRODUCTION

The absence of a generic methodology when it comes to performing text analysis has become a great challenge and a gap in research in the text mining field. This is a problem because the text mining models used are different for each case, since each area has a set of specific words with different semantic meanings. For example, the text mining model used to analyze messages on the social network Twitter is very different from the text mining model used to analyze the answers to open-ended questions in a given survey [1]. One of the most powerful and widely used techniques for text mining, the recovery of hidden information in texts and social network analysis is the topic modeling [2]. Based on these premises, it is appropriate to develop a model with generic criteria that guarantee the validity and reliability of the topic modeling that is applied to different areas. The present study emphasizes four aspects to be taken into consideration in any methodology aimed at the effective application of topic modeling: (a) a clear definition of the process of collection

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

and the pre-processing of the text database; (b) the correct selection of the parameters of the topic modeling; (c) an evaluation of the reliability of the model; and (d) a thorough interpretation of the topics identified.

On the other hand, it should be noted that the answers to open-ended questions represent an unstructured data source with very valuable information. The idea of extracting information through open-ended questions is very attractive, and widely-used in different areas and web platforms, because these data really represent the respondents' criteria [2].

The large amount of text generated by web applications and, in this particular case, surveys with open-ended questions, shows that the text type data is increasingly used. Therefore, it is necessary to deepen the study of methods aimed at automatically analyzing textual information. The application of different techniques used in natural language processing (NLP) allows the management of the human language information described in text documents [3]. According to article [4], technological development has made the traditional form of communication known as ''word of mouth'' become a new form of communication: electronic word of mouth (e-WOM). This term was coined by Goldsmith and Horowitz [5] as the internet communication that occurs through online applications and platforms such as those mentioned above and that generate a lot of textual information.

Topic modeling focuses on the grouping of text documents, assuming that each document is a function of latent variables entitled topics. In topic modeling, a topic is composed of a list of words generated through appropriate statistical methods [6]. A text can be a book chapter, blog posts, an email, answers to open-ended questions, and any type of unstructured text. Topic modeling is not aimed at understanding the semantics of words in text documents. Topic modeling is usually referred to as ''unsupervised'' methods because they involve an inference process rather than assuming the content of the topics under consideration, and have been used in a variety of fields such as social networking, software engineering, linguistic scienc, etc. [2].

Topic modeling through Latent Dirichlet Allocation (LDA), first introduced by Blei, Ng and Jordan in 2003 [6], is a computational analysis technique that can be used to investigate the thematic structure of a collection of textual data. The algorithm combines an inductive approach with statistical measures, which makes it particularly suitable for exploratory and descriptive analysis [7].

In the educational area, teacher assessment is a formal and systematic process that allows the measurement of teacher performance. The setting of teaching standards in higher education institutions requires teachers to perform effectively to meet these standards [8]. Therefore, assessing teachers in terms of identifying their virtues or shortcomings is a vitally important process. Effective teachers are expected to demonstrate high levels of teaching skills in order to meet the required standards of responsibility, and to care deeply about students and their success.

In this paper, a case study is presented in which the answers to an open-ended question contained in a teacher self-assessment survey in an Ecuadorian university are studied. The teacher assessment system implemented at the university under consideration consists of a number of components: hetero-evaluation, co-evaluation, and self-assessment. Hetero-evaluation, in the educational field, is the students' assessment of the teacher, with the aim of evaluating their performance in the teaching-learning process. The university employs a hetero-evaluation model consisting of five blocks. The first refers to the teaching methodology used by the teacher. The second covers the fulfillment of the objectives of the subject. The third allows the student to make a comparison of the course being evaluated with other subject taken. The fourth reflects the expectations of the student in this subject. Finally, the fifth allows the student to express himself openly on some additional topics related to teacher assessment. Co-evaluation or also called peer evaluation it is the observation of a class session in which a teacher of the same area of knowledge, serves as an observer for one of they colleagues. Self-assessment is the process through which teachers engage in a self-analysis of their performance in the teaching-learning process. Teachers are the best judges of their own performance, as they are able to take responsibility for much of their own professional development [8].

In the university under consideration, the teachers' self-assessment process is carried out through a online survey that contains 12 open-ended questions. Open-ended questions provide significant data in the types of surveys and questionnaires that are used in teacher self-assessment processes. Such data can provide researchers with information on the respondents' attitudes and opinions, that cannot be easily obtained from closed-question data [9]. However, the use of open-ended questions has an associated set of analytical problems, particularly in terms of identifying coherent topics that are compatible with the questions raised [10].

This study proposes the application of a methodology to execute a topic modeling algorithm in surveys with open-ended questions, with the aim of providing information on the strategies used by teachers to improve student retention in the university.

Most organizations now use online open-ended question surveys to request feedback from stakeholders on a wide variety of topics (for example, ''how can we improve our service?''). Open-ended questions become a key component of surveys. They are used to identify opinions and clarify ambiguities that researchers have not thought before [1]. However, for large samples the task of analyzing this information manually is practically impossible. Due to a large amount of textual data that is generated in virtual environments, text analysis is becoming a rapidly growing field [11]. In this context, the topic modeling technique is a powerful tool for analyzing large amounts of textual information. However, the existence of a generic methodology that guides the application of topic modeling to evaluate answers to open-ended

questions in different contexts has not been identified in the analyzed literature. Consequently, in this article, we propose the application of a generic methodology based on the modeling of topics and the modeling of text networks, which allows researchers to gather valuable information from surveys using open-ended questions. Therefore, the present study covers the research gap identified in the literature,

Another contribution of the proposed methodology is the inclusion of text network modeling algorithms that, together with the contributions of the LDA topic modeling algorithm, provide relevant results for the proposed case study. Topic modeling tends to focus solely on the frequency of terms, while the analysis of the text network takes into account both the structure of the text and the sequence of the words used [12].

This paper is organized as follows: In Section II, we propose a review of some of the works related to the application of topic modeling for the analysis of open-ended questions. Section III describes the methodology used to analyze the textual data of the case study under consideration. In Section IV, the results of the study are presented in detail. Finally, Section V details the conclusions of the work carried out, with the aim of generating discussion points for future work.

II. LITERATURE REVIEW

In this section we describe some of the works that have focused on the use of topic modeling and text mining in order to analyze surveys involving open-ended questions.

In [11], the authors presented an analysis of topic modeling in the field of software engineering in order to specify the extent to which topic modeling had been applied to one or more software repositories. They focused on 167 articles written between December 1999 and December 2014 and evaluated the use of topic modeling in the area of software engineering. They identified and demonstrated research trends in mining unstructured repositories through topical modeling. They found that most studies focused only on a limited number of software engineering tasks.

In the field of medicine, topic modeling has been implemented in a wide variety of applications. In [13] a study of the use of topic modeling is presented in which the aim is to infer the information needs of diabetic patients based on their electronic medical records, for the purpose of recommending relevant education material. In this study, the toolkit Machine Learning for Language Toolkit (MALLET) was used, which is an integrated collection of Java code that is useful for natural language processing. The method proposed in the research, based on topic modeling, can help to select educational material relevant to a given patient. The modeling of topics with the use of MALLET was carried out with a number of topics established by the authors; however, it is suggested that to explore in a more systematic way if it is necessary to determine an optimal number of topics. Finally, the study mentions that with regard to topic modeling, it is essential to carry out the pre-processing phase of data in detail [13].

The study proposed in [2] presents an alternative, semi-automatic approach involving structural topic modeling (STM). This method incorporates specific information from the documents under consideration, such as the author's gender or political affiliation. STM has been used in a number of branches of knowledge such as political science. This article focuses on analyzing how STM is useful for researchers working with surveys that include a large number of open-ended questions. Several experiments and an analysis of the open data available in the American National Election Study (ANES) were used for the study. The model proposed in the study allows the researcher to discover topics from the data, rather than assume them. When the identified topics do not correspond to theoretical perspectives, researchers either (a) revise the model for future use or (b) retain the model and use human coding procedures. This makes the model replicable in a variety of areas.

In [14] the emphasis of the research is placed on the different approaches to text analysis, taking into consideration the different areas of academia such as communication studies, sociology and, to some extent, administrative and political science. The degree of automation versus human coding in the analysis process can vary significantly between fields, with some fields tending to favor certain techniques over others. The article contributes by providing a definition of a simplified taxonomy of existing techniques for text analysis, as well as the description of a specific guide for computer-assisted text analysis for organizational science.

In the work developed by Gurgacan, et al. [15] in the area of software engineering, a semi-automatic methodology for the modeling of topics based on LDA is applied. In this work a site for finding jobs online that offers complete search options, was used as a data source. The researchers worked with the textual contents of the job advertisements published on the site that were related to big data software engineering. The aim of the work was to identify the skill sets and knowledge domains necessary for big data software engineering. As a result of the work, a taxonomy was developed that includes the domains of essential knowledge, skills and tools needed for big data software engineering.

In the educational field, the use of text mining has not been fully exploited. The study proposed by Erkens et al. [16] proposes the application of text mining through the development of an automatic tool (Grouping and Representing Tool) oriented to the analysis and visualization of cognitive information that allows to improve the collaborative learning in the classroom. The article presents a comparison of the LDA and Vector Space Model methods for the development of the tool which has been validated in an experimental case study. As a result of the study, a significant effect of the discussion on student learning was observed. Furthermore, it was shown that this effect is especially strong when students use the developed tool during their discussion.

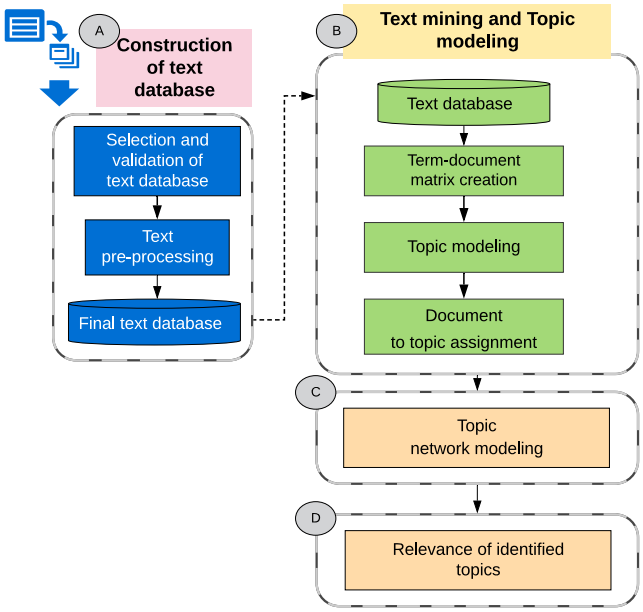


FIGURE 1. Methodology of Topic modeling and Topics network.

III. METHODOLOGY

This section describes in detail the methodology used to apply topic modeling and network modeling to a set of unstructured textual data. We work on a case study in which the answers to open-ended questions incorporated into a teacher self-assessment survey are analyzed. Fig. 1 visually describes the methodology as follows: In the first phase (A) the steps needed to collect and validate the text database are described. This means that once the textual information is collected, it is necessary to carry out a random validation process of these data in order to verify that the information collected contributes to the objectives of the study. (B) In the first instance in this phase, the text mining process is executed on the data collected and validated in phase (A). With this process, we seek to identify patterns and relationships that exist between the different elements of the text database, in order to discover new information that would otherwise be difficult to identify. In addition, this phase is complemented by the execution of topic modeling through the application of the LDA algorithm.

In the third phase of the methodology (C), a topic model network is applied. The purpose of this model is to represent any text as a network. In our case, each topic represents a node, and the connections represent the relationships that exist between them. The model identifies the most influential words in a text, based on the terms' co-occurrence [17]. Then in the model, data visualization techniques are applied to identify the different clusters which represent the main topics of the text, as well as the relationships between them. Finally, in the last stage (D), with the collaboration of an expert in the study area, the identification and description of the topics that have been automatically grouped by the computer is carried out. It should be emphasized that it is essential to carry out an analysis of the relevance of the topics identified in relation to the problem under consideration. This paper highlights the contribution of the topics identified to the strategies used to improve student retention in the university.

A. CONSTRUCTION OF THE TEXT DATABASE

At this stage, the collection and description of the text database with which the topic modeling process will operate is carried out. In addition, text database pre-processing is performed.

1) Selection and validation of text database
In terms of the selection and validation of the text database, in this phase the analysis of the data sources is carried out, and the collection process is planned. These actions aim to build a solid corpus on which the text mining and topic modeling process can be optimally executed.

2) Text pre-processing
The pre-processing stage is an important operation that guarantees the quality of the data, especially in the analysis of unstructured textual contents [18], [19]. In the present study, the pre-processing applied to the set of experimental data was determined through several sequential steps. The first step in the proposed model aimed to eliminate unwanted and irrelevant noise and data. The CSV file used the UTF-8 encoding process in order to recognize all the special characters of the Spanish language. In addition, all responses were converted to lowercase and the stop words, extra white spaces, punctuation and numbers were removed. In addition, to enrich the pre-processing phase, a process of stemming and lemmatization was carried out. Stemming and lemmatization are methods for the normalization of words. Stemming is a method used in NLP to reduce a word to its root or stem. In linguistics, lemmatization is the process of grouping the different flexed forms of a word so that they can be analyzed as a single element [20].
The words with the highest repetition in the corpus are put through a process of synonym analysis. With these words, a study of the key words in context was carried out in order to ensure that replacing one of these words with its synonym does not change the meaning of the sentence. In addition, the replacement of words in the plural by words in the singular was performed, such as "subjects" by "subject".

B. TEXT MINING AND TOPIC MODELING

1) Term-document matrix creation
In NLP a document is usually represented by a bag of words that is actually a term-document matrix. A word-document matrix is a simplified representation of the corpus, and it becomes the input used in topic modeling [21]. The order in which the documents enter the corpus does not imply that there is any relationship among them since, in the final term-document matrix, all documents and their terms are mixed to perform the

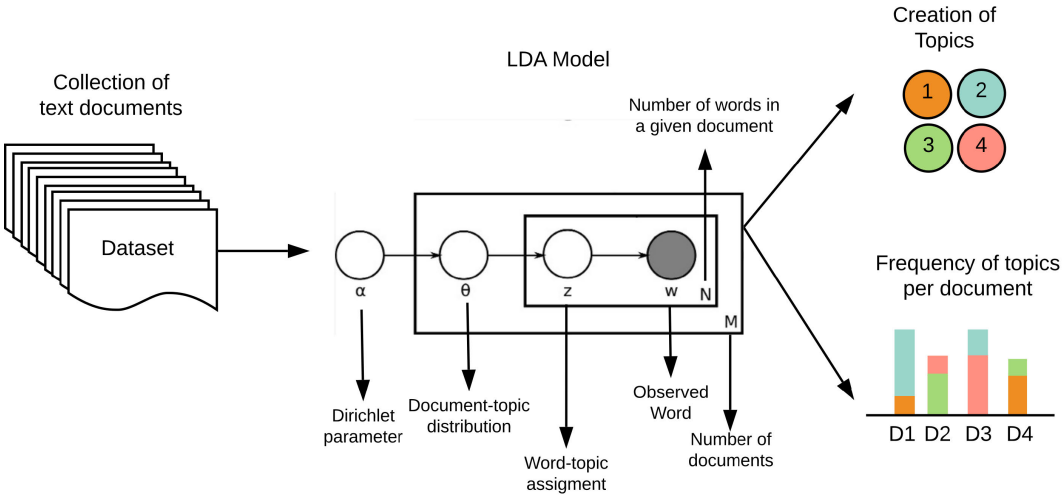


FIGURE 2. Schematic of LDA algorithm.

necessary statistical analysis. The interchangeability of words and documents could be considered the basic assumption on which the Probabilistic Latent Semantic Analysis (PLSA) and LDA topic modeling algorithms are based [22].

Given the definition of the term-document matrix, and depending on the number of documents that make up a corpus, these matrices tend to be very large. Therefore, it is common in text mining to eliminate sparse terms, that is, terms that appear in very few documents. Normally, with this process, it is possible to drastically reduce the size of the matrix without losing significant relationships inherent to it.

2) Topic modeling

This is a statistical approach to grouping text documents; it is based on the premise that each document is a function of latent variables called topics. This approach has been widely used in recent years in the field of computer science, with a specific focus on text mining and information retrieval [1]. Topic modeling methods are based on the existence of hidden variables (topics) that explain the similarities between observable variables (documents). The main and most frequently-used algorithm for modeling topics is the LDA [6]. The algorithm is a probabilistic generative model of topics, the basic idea of which is that a document is composed of a random mixture of latent topics. Each document is modeled as a mixture of bag-of-words topics, and each topic is a discrete probability distribution that defines the probability that each word appears on a given topic. LDA is considered an unsupervised generative probabilistic method for modeling a corpus. Fig. 2 shows in detail the operation of the algorithm. LDA assumes that each document (M), which is composed of a number of words (N), can be represented as a probabilistic distribution of Dirichlet on latent topics. Where α represents Dirichlet prior weight of topic by document; Z represents the assignment of a word to a given topic, and W represents the observed word in document M. In the present study, LDA is used to discover the topics that occur in the documents.

3) Document to topic assignment

This is perhaps the best advantage of the LDA algorithm. In the generative process, the algorithm assumes that a word belongs to a topic, and that a document belongs to at least one topic. Under this premise, it is necessary to correctly select the parameters of α, which is the distribution of the topics per document. If a high value of α is selected, the distribution of the topics will be homogeneous, while a low value of α prevents the inference process from distributing the percentage of probability in some subjects.

C. TOPIC NETWORK MODELING

The document to topic assignment matrix described in point 3 of phase (B) is binarized in order to select the most relevant topics per document. The resulting binary matrix relating documents to topics can be represented as a bipartite network, which can be projected [23] to analyze the topic relationships.

D. RELEVANCE OF IDENTIFIED TOPICS

The objective of this stage is to try to find a label that describes the substantive content of the set of topics, automatically grouped by the algorithm [24]. The interpretation of the resulting topics involves a synthesis process in which the contribution of an expert in the subject is a fundamental element.

IV. RESULTS OF THE CASE STUDY

Before performing the proposed methodology described in Fig. 1, a manual corpus-tagging process was carried out, in which approximately 10 % of the total responses were

TABLE 1. Topics and tokens identified in manual random analysis.

Topics	Tokens
Practical teaching	Application, practice, exercises, practical, interactive, workshops, cases, experiments, laboratory, outings, field, linking, community, visits, study, analysis, clinical.
Experiential learning	Situations, life, real, real, problems, professional, experiential, experience, personal, experiences, everyday, guests, experts, consulting.
Teaching tutorships	Customized, feedback, conversation, dialogues, leveling.
Use of technology	Video, audiovisual, technology, resources, information, classrooms, virtual, environments, platforms, digital, simulation, simulate.
Types and mechanisms of evaluation	Rubrics, evaluation, summative, systematic, periodic, continuous, permanent, tests, exams.
Group and collaborative work	Work, group, collaborative, participation, peers, couples, team.
Teaching-learning environment	Environment, treatment, pleasant, flexibility, creating, trust, atmosphere, respect, mutual, climate, cordiality, cohesion.
Personalized follow-up to the student	Follow-up, motivation, accompaniment, encourage, communication, knowledge, well-being, student, listen, empowerment, approach, personal, interest, empathic.

randomly analyzed. The process consisted of the general reading of the selected answers in order to identify macro topics to which the teachers refer in their answers. As a result of this process, 16 topics were initially identified as summarizing the teachers' responses. In a second phase of the manual analysis some topics were merged, leaving 8 topics. In addition, the most representative tokens were identified for each topic. Table 1 shows the topics and tokens that were identified in the manual labelling process. This randomized manual analysis provided valuable information in terms of establishing criteria for the process of text pre-processing (the removal of extra white space, punctuation, stop words; stemming and lemmatization).

The results obtained by applying the proposed methodology to the case study are detailed below.

A. CONSTRUCTION OF THE TEXT DATABASE

For this study, we used a text database of approximately 900 answers that correspond to the open-ended questions: ''Indicate which strategies you have adopted to improve student retention in your classes without affecting academic quality. Include specific examples regarding your strategies''. The question is such that it offers a structure that allows the direct identification of the strategies and examples used by teachers to improve student retention. This structure allowed respondents to address the question in a similar way and, therefore, use words consistently. This question is included in the teacher self-assessment survey executed at the university under study in the March - July 2019-2 semester. The text database of the teachers responses was worked in a CSV file. Below are two examples of teacher responses for the sake of explanation:

''During my teaching experience I have taken note that students are more attentive when conceptual and theoretical aspects are associated with real-life cases, and therefore I try to talk to them about their own or known experiences, linked to the topic being dealt with in class. The other mechanism is to pause to discuss their experiences or criteria on the subject of class.''

''The motivation that must be generated on the part of students in each class is essential. Visiting establishments in the industry is an opportunity for students to evaluate and understand the knowledge acquired in class, and its importance for their employment relationship.''

B. TOPIC MODELING USING LDA

After performing the manual labeling of the corpus, the first phase of the methodology used, the data preparation or preprocessing of the corpus was executed. From the corpus the document term matrix resulted in 5308 terms and 836 documents with a sparsity of 0.99. After a removal of terms with a sparseness larger than 0.99, that is the terms with a low relative frequency are removed, the resulting document term matrix had 387 and documents: 836 with 0.97 sparsity. A term frequency–inverse document frequency (TF-IDF) analysis was carried out in order to identify the important terms in the corpus. Once such terms had been identified, the top TF-IDF terms were removed as they were so prominent that they appeared in every topic modeling with high degree of importance. After the removal and inspection of the exploratory topics, the most frequent terms are depicted in Fig. 3. The terms appear to be evenly distributed among the final topics modeled.

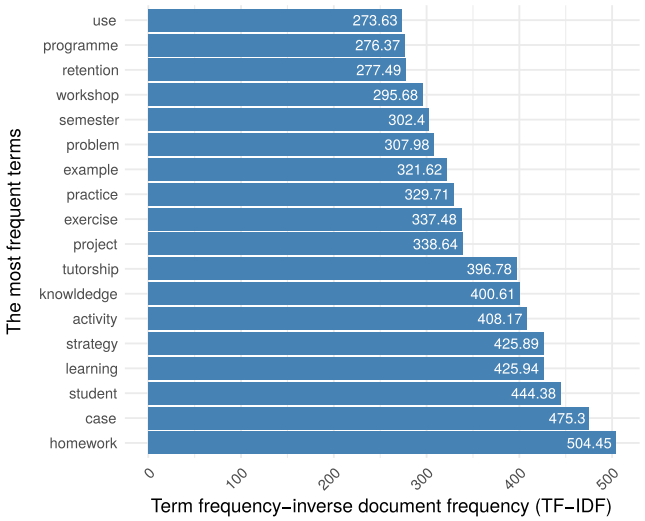


FIGURE 3. Most frequent terms in text corpus with their corresponding term frequency–inverse document frequency (TF-IDF).

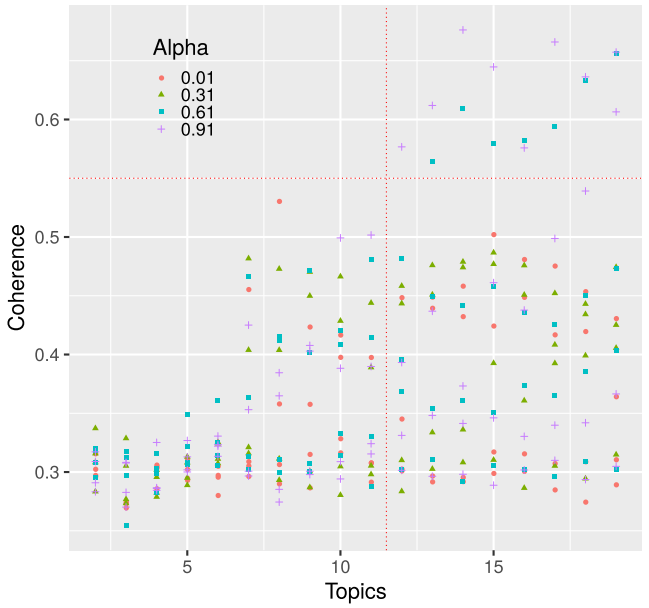


FIGURE 4. Hyper-parameters optimization for number of topics k and LDA α parameter.

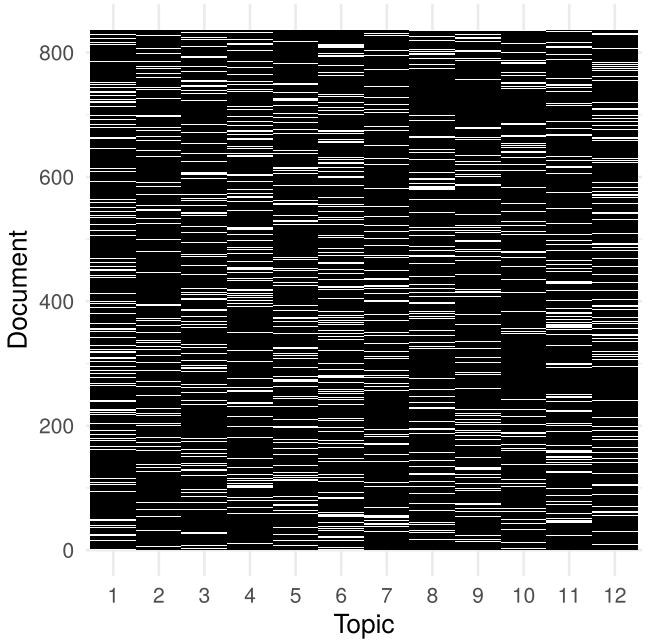


FIGURE 5. Heatmap representation of the matrix relating documents (rows) to topics (columns); white and black colors corresponds to connection (1) and no connection (0) present respectively.

After carrying out the exploratory topic modeling, an exhaustive search for the number of topics k was performed and the hyperparameter α for the LDA algorithm optimized. The number of topics from $k = 2$ to $k = 19$ were explored, and we tested the values for alpha that were uniformly distributed between 0 and 1, namely $\alpha = \{0.01, 0.31, 0.61, 0.91\}$. An evaluation measure was necessary in order to analyze the combination of parameters in an exhaustive search. To evaluate the parameters tuples (k, α) the coherence measure C_V was used. C_V is based on a sliding window, where the counts are used to calculate the pointwise mutual information of every top word to every other top word, resulting in a set of vectors, one for every top word. From this, the cosinus similarity between every top word vector and the sum of all top word vectors is calculated. The coherence is the arithmetic mean of these similarities [25]. The topics are considered to be coherent if all or most of the words, for example, the topics' top N words, are related. Here, the challenge is to obtain a measure that correlates highly with manual topic tagging to help the interpretability by humans [26], [27].

Fig. 4, depicts the results of the aforementioned exhaustive search.A region of interest can be seen in the top right corner, which represents the best parameter combinations according to the coherence measure.

This is for $k \geq 12$ and both $\alpha = 0.61$ and $\alpha = 0.91$. Although the maximum value for the coherence occurs for $k = 14$ and $\alpha = 0.91$, after inspecting the model with such parameters, a final LDA model was selected for $k = 12$ and $\alpha = 0.61$, given that more topics result in higher difficulty when it comes to finding a meaning from a human perspective, and a larger value of alpha will increase the document-to-topic assignments. That is, a higher alpha value will lead to documents being more similar in terms of what topics they contain [28]. Thus we select the lower values for k and α in the region of interest (the top-right corner), in order to have more document-to-topic diversity and a less densely-connected document-topic network. In addition, the value for the value for k is consistent with the manual corpus tagging process.

The LDA models each document as a mixture of topics. That is, each document has a probability of belonging to each topic. This can be interpreted as a bipartite graph in which each document is connected to a given number of topics. Such relationships are explored in the next subsection.

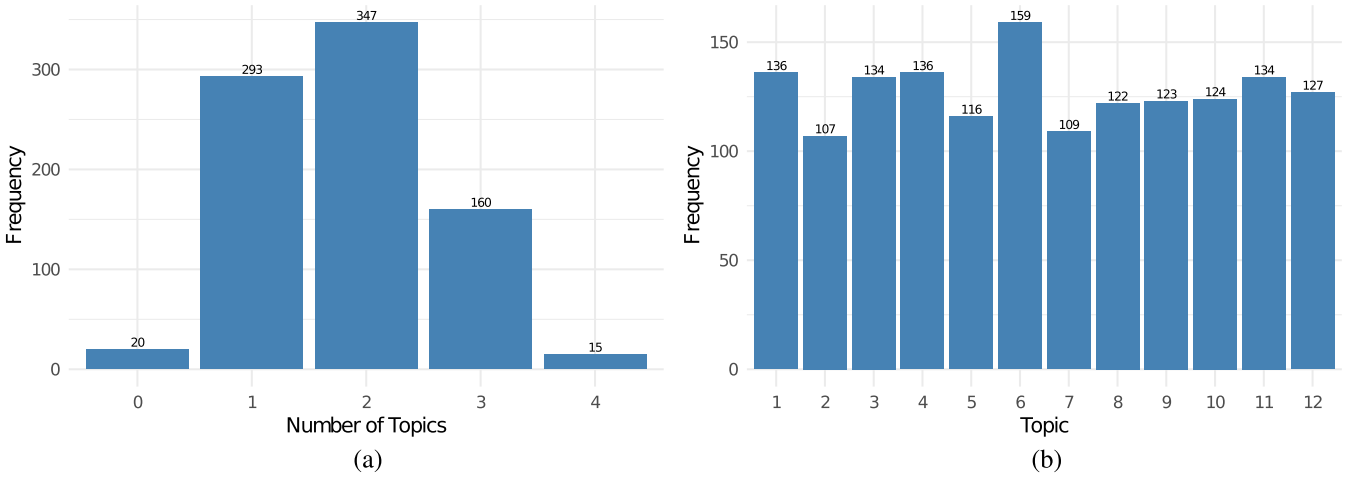


FIGURE 6. (a) Frequency of number of topics assigned to documents. (b) Topics assignment frequencies.

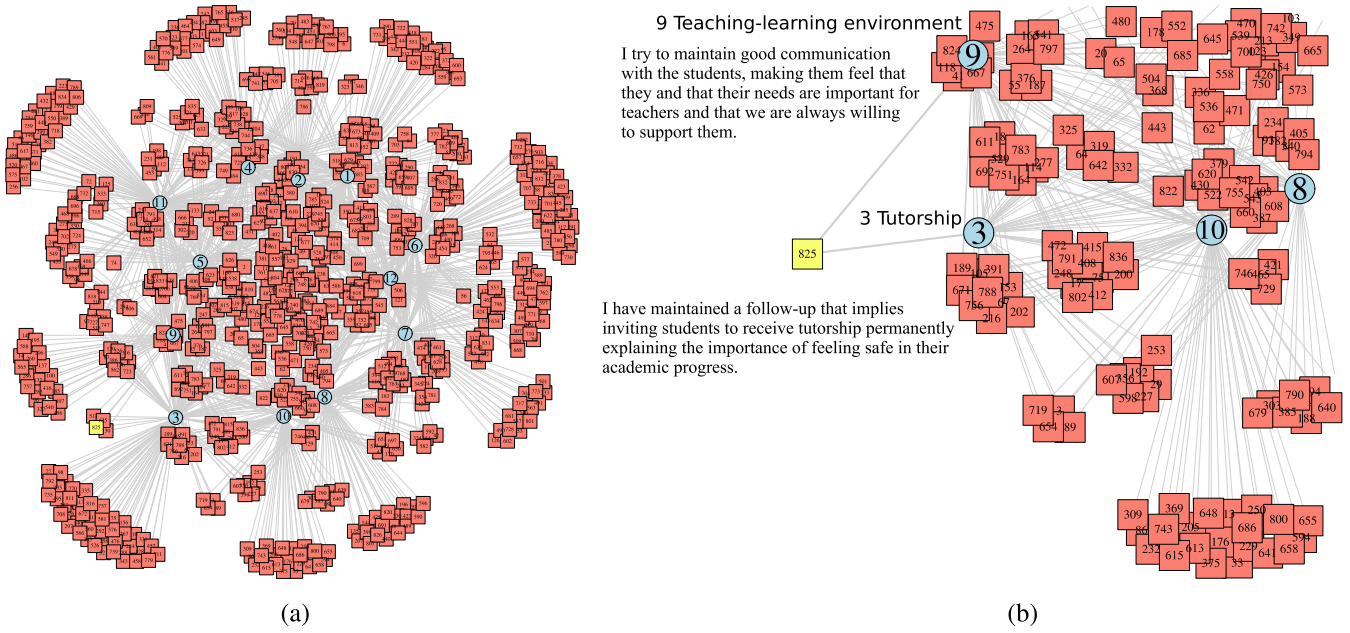


FIGURE 7. (a) Bipartite network representation connecting documents (red squares) to topics (blue circles). (b) Zoomed portion of the network, including a teacher's response (825) and the interrelation with topics 3 and 9.

C. TOPIC NETWORK MODELING

From the probability per document-topic matrix obtained in the last subsection through topic modeling using LDA, a network relating documents to topics is constructed. A binarization threshold of $\theta = 0.15$ is used to remove/create relations between document and topics. The original continuous probability matrix $\mathbf{P}$ is binarized as follows: if $P_{ij} < \theta$, $P^B_{ij} = 0$, otherwise $P^B_{ij} = 1$, where $\mathbf{P^B}$ is the resulting binary matrix relating documents to topics. Fig. 5 depicts a heatmap of the probability per document-topic matrix relating documents (rows) with topics (columns), as defined by matrix $\mathbf{P^B}$, where the black and white colors correspond to 0 and 1, i.e. a connection is present or not present, respectively. Note that the sum of the columns will give the topic frequencies as shown in Fig. 6 (a). The number of topics assigned to each document is depicted in Fig. 6 (b). This results from the rows' sum, after summarizing the sum frequencies.

Fig. 7 (a) represents the document-per-topic matrix depicted in the left panel as a bipartite network, where the topics are represented as pastel blue circles, and the documents as pastel red squares. Fig. 7 (b) depicts the zoomed bottom-left part of the network, where one can appreciate the connections occurring between documents (red squares) and topics (blue circles), but not between nodes of the same type, as it is a bipartite network. A yellow square corresponding to an example of a teacher's response (825), is highlighted,

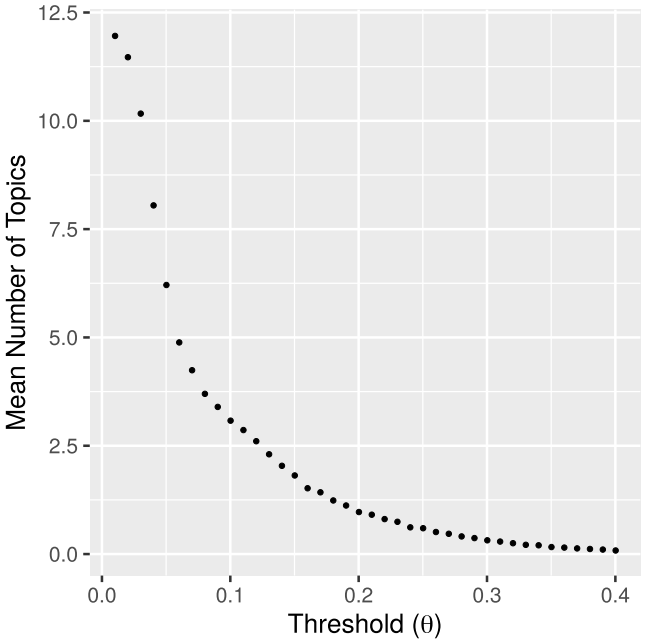


FIGURE 8. Mean number of topics assigned to documents according the binarization threshold θ.

connected with topics 3, tutorship, and the associated related text: "I have maintained a follow-up that implies inviting students to receive tutorship, permanently explaining the importance of feeling safe in their academic progress." Also, a connection with topic 9, the teaching-learning environment, is presented with the related text: "I try to maintain good communications with the students, making them feel that they and that their needs are important for teachers, and that we are always willing to support them."; belonging to the same teacher response (825).

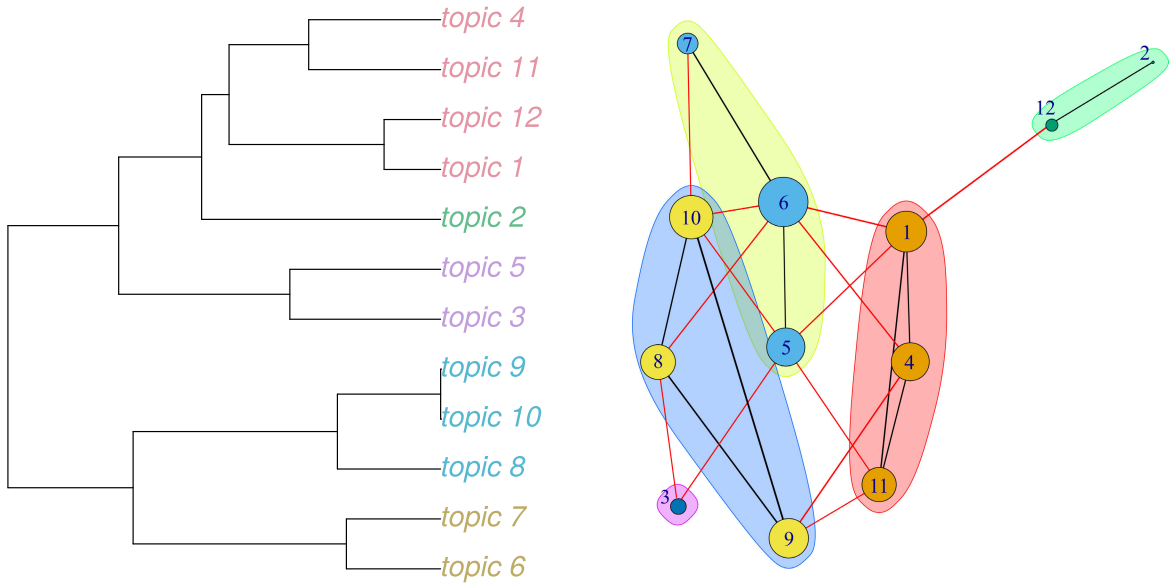
The binarization threshold $\theta = 0.15$ was selected empirically using the mean number of topics assigned to each document. An exhaustive search for values of the binarization threshold from $\theta = 0.01$ to $\theta = 0.4$, was carried out and the mean value of the number of topics assigned to documents was calculated for each threshold value and used to decide the value of θ. The results can be appreciated in Fig. 8, where an elbow behavior occurs between $\theta = 0.1$ and $\theta = 0.2$. Thus the value of $\theta = 0.15$ is selected as the final binarization threshold used above. For instance for Fig. 6 (a), the mean number of topics is 1.82 for $\theta = 0.15$. Selecting a higher value for θ will end up with more documents disconnected in the network (approximately 0.1 as the mean number of topics assigned). A lower value of θ will end up with the majority of documents assigned to all 12 topics (mean number of topics assigned near 12), which is equivalent to a more densely connected network. The frequencies for the selected $\theta = 0.15$ are depicted in Fig. 6 (a).

The most prevalent are 1, 2 and 3 topic assignments with frequencies of 293, 347 and 160 respectively. Only 20 documents were assigned to no topic (0 number of topics) being disconnected and removed from the bipartite network in Fig. 7 (a), while 15 documents were assigned to 4 topics. On the other hand, Fig. 6 (b) shows the frequency per topic, which is uniform. That is, the 12 topics were assigned uniformly to the documents. Table 2 shows the most common tokens for each topic.

A bipartite network projection [23] is carried out to discover the relationship between the topics identified. The projection relates topics that are common to documents, and the number of documents in common giving the weight between topics, giving a measure of how strong their relationship is. Then, the weights are min-max scaled between 0 and 1. A distance measure d_{ij} between topics i and j can be written as the complement of the scaled weights: $d_{ij} = 1 - \text{scaled\_weight}$. The relationships between topics are then depicted in Fig. 9 (a) as a dendrogram taking into account the afore-mentioned distance. Fig. 9 (b) depict the relationships between topics as a graph, where the communities' structure has been identified. Communities reveal how a network is internally organized, and indicate the presence of special relationships between the elements of a system [29], in this case the relationship between topics. The selected community detection algorithm is based on edge betweenness. This measures the number of shortest paths through a given edge. Edges connecting separate modules have high edge betweenness, as all the shortest paths from one module to another must pass through them. Such edges are removed to create a hierarchical map, called a dendrogram of the graph [30]. Such a structure is depicted in Fig. 9 (b) for the topics graph. Another projection of the bipartite network is possible [23] relating the respondents (documents). Such a projection is not shown here. Given the anonymity of the survey, no additional information would be present in the respondents' network to allow for an additional analysis. In the next sub-section we discuss the topics' relevance and relationships.

D. RELEVANCE OF THE IDENTIFIED TOPICS

In the topic modeling process, human interaction at the beginning and at the end of the stages of the methodology plays a crucial role in determining the quality of the final model. The identification of topics based exclusively on statistical tools and computational processes causes the semantic structure of documents to be lost. Therefore, the topics identified by the algorithm do not necessarily describe the content of a document. At the beginning of this project, a manual reading process of approximately 10 % of teachers' responses was carried out. This process allowed a first identification of the main topics which teachers mention when filling out the survey. It is worth noting that the objective of applying the proposed methodology to the case study is to synthesize the teachers' answers to the question "Indicate which strategies you have adopted to improve student retention in your classes without affecting academic quality. Include specific examples regarding your strategies". In other words,



(a) (b)

FIGURE 9. Clustering of topics network. (a) Dendrogram with 5 clusters using the scaled weights complement as distance. (b) Topics graph and community structure using edge betweeness community detection algorithm.

TABLE 2. Word classification by topic.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	perform	learning	tracing	use	work	case	knowledge	strategy	always	professional	semester	work
2	questions	project	tutoring	virtual	case study	exercise	internship	student	have	life	final	activity
3	workshop	process	performance	tool	student	practical	application	retention	treatment	strategy	progress	tutoring
4	reading	evaluation	trouble	videos	groups	analysis	methodology	interest	importance	career	tutoring	group
5	presentation	teacher	academic	participation	different	real	makes	to get better	student	trouble	exams	group
6	content	career	tutoring	activity	same	concepts	beef up	example	teacher	student	evaluation	individual
7	readings	part	difficulties	technology	climate	examples	perform	level	activity	real	grade	practice
8	debates	teaching	tasks	through	developing	clinical	part	quality	case	personal	to get better	custom
9	discussion	feedback	evaluations	material	first	workshop	permitted	academic	trust	give	grades	student
10	base	evaluations	personal	workshop	various	games	higher	principal	they can	case	exam	following
11	participation	results	low	information	ability	study	perform	may	semester	examples	tests	account
12	investigation	elaboration	hours	active	technique	realization	theoretical	times	ambient	theory	recovery	doubts
13	related	needs	constant	communication	semester	dynamics	patients	inside	support for	know	present	teams
14	through	example	dedication	inside	method	field	simulation	question	motivation	experiences	shape	additional
15	essays	understanding	permanent	dedication	investigation	explanation	new	generate	important	I give	week	advance
16	exhibitions	virtual	student	exercise	thought	project	apply	achieve	attention	world	tasks	results
17	controls	design	tutorship	equipment	two	studies	student	face-to-face	help	chats	companions	It allows
18	then	peers	to get better	programming	foment	understand	doing	end	time	real	additional	presentation
19	discussions	autonomous	identify	support for	utilization	own	theoretical	many	start	additionally	summary	evaluate
20	information	based	custom	technique	teacher	advances	theorists	concept	keep	then	task	mistakes

this study aims, through a semi-automatic model, to identify the main strategies that teachers use to improve student retention in the university. Table 2 presents the 12 relevant topics of the case study which have been identified after applying the LDA algorithm to the analyzed corpus. In the present work, with the participation of an expert, we worked on the identification of topics that describe the substantive content of each set of topics generated by the algorithm. Different colors were assigned to identify the tokens that contribute a semantic meaning to each of the topics described in Table 3.

For a better analysis of the topics identified and described in Table 3, the topics have been grouped according to the clusters observed in the graph in Fig. 9 (b). The classification is as follows: Cluster (yellow) Practical Learning – includes the topics 5, 6 and 7; Cluster (blue) Teaching Initiative - includes the topics 8, 9 and 10; Cluster (red) Use of Technological Tools and Traditional Teaching Strategies – includes the topics 1, 4, and 11, and Cluster (green) where cluster 12 contains tokens related to the Teamwork Strategy.

TABLE 3. Topic identification by token group.

Topic	Description
Topic 1	Research, analysis and reading
Topic 2	No definition
Topic 3	Tutorship
Topic 4	Use of technology
Topic 5	No definition
Topic 6	Practical learning
Topic 7	Practical learning
Topic 8	Retention strategies
Topic 9	Teaching-learning environment
Topic 10	Experiential learning strategy
Topic 11	Evaluation mechanisms
Topic 12	Team work

On the other hand, a specific topic has not been identified for cluster 2.

Topics 6, 5 and 7 refer to the use of strategies that promote practical learning as the mechanism most used by teachers to support student retention. In the Fig. 9 (a) and the Fig. 9 (b) the prevalence of topics 6 and 7 can be observed, as well as the weight of their relationships in the case study. Practical learning is evidenced through the development of workshops, case studies, use of simulators, exercises linked to a central topic, gamification and dynamics for interactive work. On the other hand, it can be seen in the dendrogram of Fig. 9 (a) that topic 5 contributes to this topic in a very slight way, which is corroborated by the manual classification of the text.

The terms immersed in topic 10 refer to experiential learning strategies, that is, the inclusion of experiences of the professional field to motivate student learning. In the graph of Fig. 9 (a) a particular relationship of topic 10 with topics 5, 6 and 7 can be observed. This makes sense since there is a very close relationship between practical learning and experiential learning. In this same cluster are topics 8 (student academic retention) and 9 (teaching-learning environment). Topic 9 (teaching-learning environment) groups terms related to the empathic relationship that must exist between teachers and students, a relationship that generates confidence and motivation on the part of the student. It is interesting to note that the three topics (8, 9 and 10) relate to group strategies, which are not necessarily part of a curriculum, but are particular initiatives on the part of some teachers which contribute to the student-teacher relationship and therefore improve the teaching-learning process and academic retention.

Topics 1 (Reading, analysis, and research), 4 (Use of technology) and 11 (Evaluation mechanisms) are part of a new cluster. Topic 4 is at the center of the cluster, which makes us think that the use of technology in the classroom should be a central element when it comes to capturing the attention of students. On the other hand, topics 1 and 11 should be seen as tools to strengthen the activities programmed in virtual environments.

Special attention deserves to be paid to topic 3, since terms are grouped in it that are related to the execution of academic tutoring. In the academic field, this activity is essential to strengthen student learning and therefore enhance their permanence (retention) in the university. In the graph of Fig. 5(a) this analysis is corroborated, since a direct relationship is observed between topic 3 and topic 8 (academic retention of students).

Topic 2, as can be seen in the dendrogram and the graph in Fig. 9 (a), is the one that has the least relationship with other topics, and therefore has the least relevance in the study. This is corroborated by the manual analysis, since the terms included in this topic are much dispersed and do not point to a specific topic.

V. CONCLUSION

In university settings, surveys are constantly carried out involving teachers, students, graduates and employers. These surveys have the function of collecting valuable information with the objective of identifying the degree of satisfaction that the above-mentioned actors have with the academic processes carried out in a university. These surveys include open-ended questions with the aim of detecting spontaneous thoughts and exploring the attitudes of the respondents. However, open-ended questions also have a great challenge: their analysis is associated with a large workload and time. This often avoids including open-ended questions in the surveys, thus losing the opportunity to gather valuable information from those involved in a process. Here lies the importance of the present study, in which we propose the application of a generic methodology based on the modeling of topics and the modeling of text networks, which allows researchers to gather valuable information from surveys with open-ended questions. The application of this methodology will allow optimizing the work and time necessary to comply with the analysis of the textual information generated by the open questions. It should be noted that the proposed methodology is not only specific to the educational area, but can also be replicated to other areas, such as those described in the present research.

The present work describes from the beginning to the end the methodology involved in extracting hidden information from the textual data of a survey through topic modeling. A differentiating aspect of this article from other similar works is the application of text network modeling as a complementary tool to topic modeling, in order to strengthen the text analysis of open-ended questions. The topic modeling case study described in this paper is designed for the analysis of answers to open-ended questions in the field of teacher self-assessment. However, with minor modifications, the model is flexible enough to be used with other unstructured data sources.

The structure of the proposed question associated with the case study offers an answer focused on a particular requirement of teacher self-assessment (“...identify strategies to improve student retention”). This structure allowed all teachers to address the question in a similar way and, therefore, use words consistently. This consideration is fundamental aspect that should be incorporated when planning to apply topic modeling to open-ended questions.

In the present study, limitations were identified such as a fairly limited sample of data. The extraction of topics in short texts becomes difficult because the traditional methods and algorithms of topic modeling are based on the word co-occurrence identified in a text. However, such co-occurrence is not common in short texts, which means that conventional algorithms suffer from a severe data shortage. As for the guidelines to take into account in future research, it would be desirable to work with a corpus with more data.

Therefore it is suggested that there is a need to enrich the pre-processing phase with the application of the following techniques: disambiguation and part of speech tagging, entity extraction (recognize proper names such as locations, organizations, or names of people) and n-gram detection (identify words that are grouped into a single term). On the other hand, another limitation to consider is that the manual analysis of the topics identified by the algorithm, as is done in this case study, is influenced by the perceptions and background of the researcher.

In section Results of the case study, in section D (Relevance of the identified topics) a detailed analysis of the topics identified is presented after having applied step by step the proposed stages in the methodology. In this description it can be observed that, in the proposed case study, the topics obtained allowed to clearly identify the main strategies that teachers have applied to improve student retention. This allows us to conclude that the proposed methodology can be applied in different fields to analyze surveys with open-ended questions.

Based on the results of the case study developed in this article, it is suggested that future research should aim to identify additional variables that categorize the respondents, for example in terms of age, gender, specialty, temporary dedication, educational level, among others. These parameters would allow a better clustering of topics and, therefore, more specific results would be obtained.

REFERENCES

[1] A.-S. Pietsch and S. Lessmann, "Topic modeling for analyzing open-ended survey responses," *J. Bus. Anal.*, vol. 1, no. 2, pp. 93–116, Apr. 2019, doi: 10.1080/2573234X.2019.1590131.

[2] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural topic models for open-ended survey responses," *Amer. J. Political Sci.*, vol. 58, no. 4, pp. 1064–1082, Mar. 2014.

[3] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Nov. 2018.

[4] A. Reyes-Menendez, J. R. Saura, and J. G. Martinez-Navalon, "The impact of e-WOM on hotels management reputation: Exploring TripAdvisor review credibility with the ELM model," *IEEE Access*, vol. 7, pp. 68868–68877, 2019.

[5] R. E. Goldsmith and D. Horowitz, "Measuring motivations for online opinion seeking," *J. Interact. Advertising*, vol. 6, no. 2, pp. 2–14, Mar. 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[7] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam, "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology," *Commun. Methods Measures*, vol. 12, nos. 2–3, pp. 93–118, Feb. 2018, doi: 10.1080/19312458.2018.1430754.

[8] M. Akram and S. Zepeda, "Development and validation of a teacher self-assessment instrument," *Res. Reflections Edu.*, vol. 9, no. 2, pp. 134–148, 2015.

[9] J. A. Ross and C. D. Bruce, "Teacher self-assessment: A mechanism for facilitating professional growth," *Teach. Teacher Edu.*, vol. 23, no. 2, pp. 146–159, Apr. 2007.

[10] W. H. Finch, M. E. H. Finch, C. E. McIntosh, and C. Braun, "The use of topic modeling with latent Dirichlet analysis with open-ended survey items," *Transl. Issues Psychol. Sci.*, vol. 4, no. 4, pp. 403–424, Dec. 2018.

[11] T.-H. Chen, S. W. Thomas, and A. E. Hassan, "A survey on the use of topic models when mining software repositories," *Empirical Softw. Eng.*, vol. 21, no. 5, pp. 1843–1919, Sep. 2015.

[12] K. Lee, H. Jung, and M. Song, "Subject–method topic network analysis in communication studies," *Scientometrics*, vol. 109, no. 3, pp. 1761–1787, Sep. 2016.

[13] S. Kandula, D. Curtis, B. Hill, and Q. Zeng-Treitler, "Use of topic modeling for recommending relevant education material to diabetic patients," in *Proc. AMIA Annu. Symp.*, 2011, p. 674.

[14] G. C. Banks, H. M. Woznyj, R. S. Wesslen, and R. L. Ross, "A review of best practice recommendations for text analysis in r (and a user-friendly App)," *J. Bus. Psychol.*, vol. 33, no. 4, pp. 445–459, Jan. 2018.

[15] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling," *IEEE Access*, vol. 7, pp. 82541–82552, 2019.

[16] M. Erkens, D. Bodemer, and H. U. Hoppe, "Improving collaborative learning in the classroom: Text mining based grouping and representing," *Int. J. Comput.-Supported Collaborative Learn.*, vol. 11, no. 4, pp. 387–415, Nov. 2016, doi: 10.1007/s11412-016-9243-5.

[17] D. Paranyushkin, "InfraNodus: Generating insight using text network analysis," in *Proc. World Wide Web Conf.*, 2019, pp. 3584–3589.

[18] A. Kyriakopoulou and T. Kalamboukis, "The impact of semi-supervised clustering on text classification," in *Proc. 17th Panhellenic Conf. Informat. (PCI)*, 2013, pp. 180–187.

[19] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, NJ, USA: Wiley, 2011.

[20] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proc. 13th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2004, pp. 625–633.

[21] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, Sep. 2016.

[22] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Inf. Retr.*, vol. 14, no. 2, pp. 178–203, Aug. 2010.

[23] D. B. Larremore, A. Clauset, and A. Z. Jacobs, "Efficiently inferring community structure in bipartite networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 1, Jul. 2014, Art. no. 012805.

[24] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digit. Journalism*, vol. 4, no. 1, pp. 89–106, 2016.

[25] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2015, pp. 399–408.

[26] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2017, pp. 165–174.

[27] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5645–5657, Aug. 2015.

[28] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," *Document numérique*, vol. 17, no. 1, pp. 61–84, Apr. 2014.

[29] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 1, Jul. 2009, Art. no. 016118.

[30] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.

DIEGO BUENAÑO-FERNANDEZ received the engineering degree in computer systems from the National Polytechnic School, Quito, in 1999, and the master's degree in business administration from Latin American Christian University, in 2012. He is currently pursuing the Ph.D. degree, in the Ph.D. program, in computer science with the University of Alicante, Spain. He is also the Dean of the Faculty of Engineering and Applied Sciences, Universidad de Las Américas, Quito, Ecuador. He also teaches the subjects operating systems and electronic business. His research line is related to data mining in educational environments.

MARIO GONZÁLEZ received the Ph.D. degree in computer science from the Autonomous University of Madrid (UAM), in 2012. He is specialized in the area of artificial intelligence, complex systems, and information processing using neural networks. His research includes the modeling of attractor networks for pattern retrieval and data analytics.

DAVID GIL is currently an Assistant Teacher 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 articles. His main research topics include artificial intelligence applications, data mining, open data, big data, and decision support systems in medical and cognitive sciences.

SERGIO LUJÁN-MORA received the Ph.D. degree in computer engineering from the Department of Software and Computing Systems, University of Alicante, in Spain, in 2005, and the degree in computer science and engineering from the University of Alicante, in 1998. He is currently a Senior Lecturer with the Department of Software and Computing Systems, University of Alicante. In recent years, he has focused on e-learning, massive open online courses (MOOCs), open educational resources (OERs), and the accessibility of video games. He has authored several books, many published articles in various conferences, including ER, UML, and DOLAP, and high-impact journals, including DKE, JCIS, JDBM, JECR, JIS, JWE, IJEE, and UAIS. His main research interests include web applications and web development, and web accessibility and usability.