IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 22, 2021, accepted May 16, 2021, date of publication May 24, 2021, date of current version June 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083000

TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition

CIPRIAN-OCTAVIAN TRUICĂ AND ELENA-SIMONA APOSTOL
Department of Computer Science and Engineering, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, 060042 Bucharest, Romania

Corresponding author: Elena-Simona Apostol (elena.apostol@upb.ro)

This work was supported by the University Politehnica of Bucharest through the PubArt Program.

ABSTRACT Topic modeling is a probabilistic graphical model for discovering latent topics in text corpora by using multinomial distributions of topics over words. Topic labeling is used to assign meaningful labels for the discovered topics. In this paper, we present a new topic labeling method that uses automatic term recognition to discover and assign relevant labels for each topic, i.e., TLATR (Topic Labeling using Automatic Term Recognition). TLATR uses domain-specific multi-terms that appear in the set of documents belonging to a topic. The multi-term having the highest score as determined by the automatic term recognition algorithm is chosen as the label for that topic. To evaluate TLATR, we use two real, publicly available datasets that contain scientific articles' abstracts. The topic label evaluation is done both automatically and using human annotators. For the automatic evaluation, we use Pointwise Mutual Information, Normalized Pointwise Mutual Information, and document similarity. For human evaluation, we employ the average rating method. Furthermore, we also evaluate the quality of the topic models using the Adjusted Rand Index. To prove that our novel method extracts relevant topic labels, we compare TLATR with two state-of-the-art methods, one supervised and one unsupervised, provided by the NETL Automatic Topic Labelling system. The experimental results show that our method outperforms or provides similar results with both NETL's supervised and unsupervised approaches.

INDEX TERMS Automatic term recognition, automatic topic labeling evaluation, topic labeling, topic modeling.

I. INTRODUCTION

As the volume of textual data grows daily, it becomes increasingly important to use appropriate models and methods for improving the knowledge extraction process. In text mining and natural language processing, topic modeling is a statistical model used to discover the hidden semantic structures of words from textual data [10]. To better understand the topics, topic labeling is used to quantify in a human readable way the semantic meaningfulness of a latent topic [39].

In this paper, we focus on automatic topic labeling, a research field that has seen a lot of attention in recent years. We propose a new method, i.e., TLATR (Topic Labeling using Automatic Term Recognition), that uses automatic (domain-specific) term recognition (ATR) for determining the terms that are relevant to a topic. C-Value is used for extracting and scoring the domain-specific terms [15]–[17]. The topic label is chosen as the domain-specific term with the highest C-Value score. Based on previous results presented in [54] and [52], we use two topic modeling algorithms and two weighting schemes. The topic modeling algorithms are: Latent Dirichlet Allocation (LDA) [6] and Non-negative Matrix Factorization (NMF) [11], [23]. The term weighting schemes are *TF-IDF* [41] and *Okapi BM25* [49], [50]. Moreover, by employing linguistic filters for selecting candidate terms, we add a new context dimension to topic labeling, a dimension that, to the best of our knowledge, is lacking in the current literature.

TLATR consists of two main steps. Firstly, it incorporates the context dimension to determine a list of representative domain-specific candidate terms. These candidate terms are extracted from the documents belonging to a topic using linguistic filters. Secondly, we extract the labels for each topic from the list of candidate terms. As these topic labels

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

are extracted from the list of representative domain-specific candidate terms, they are also representative for the topic and the set of documents belonging to a topic. Thus, they discover new hidden semantic structures in a text corpus, i.e., terms that are semantically similar while syntactically different. Semantically similar terms are synonyms that have the same meaning and can be used in the same context. Our method is strictly unsupervised, i.e., there is no labeled corpus from which to build the model. Everything is learned dynamically, no model is built and the labels are highly dependent on the quality of the topics and the documents belonging to these topics.

We collected two publicly available, already labeled, corpora containing articles from the scientific domain. The number of labels for each corpus is used to initialize the number of topics to be discovered by the topic modeling algorithm. The two corpora contain documents that are not totally disjoint, meaning that the concepts presented in the documents overlap. Furthermore, we propose the use of these corpora as new datasets that could facilitate the evaluation of the topic labeling task.

We utilize Pointwise Mutual Information (PMI) and Normalized Pointwise Mutual Information (NPMI) to evaluate the labels extracted by the proposed method, as they showed promising results [39]. As is done in clustering evaluation [9], [35], a topic label with a high PMI or an NPMI closer to 1 is relevant to the documents belonging to that topic, therefore it is a good label for the topic. Furthermore, this type of evaluation underlines how representative are the labels to the topic and to the set of documents belonging to the topic. We also evaluate the topic model results using the Adjusted Rand Index.

To evaluate TLATR, we also employ 30 human annotators and we compute the Average Rating. Measures, such as nDCG and Top-1 average rating, are not used for evaluating the quality of the labels detected by our method as done in [2], [5], [25], as we do not ask the human judges to rank multiple labels for the same topic but only to score how appropriate is the label for the topic. However, we firmly believe that human evaluation should not be used because they suffer inconsistencies [62], and, in some cases, the inter-annotator F-measure between human annotators is $70 - 80\%$ [22].

To evaluate the similarity between the topic label and the topics, we use word embedding to represent both the topic label and its keywords as word vectors. The similarity score is given by an average cosine similarity computed for the word vectors.

We compare our proposed method with a state-of-the-art solution, NETL [5], that has both a supervised and unsupervised approach, and its code is freely available online. Furthermore, we could not consider for comparison the majority of the existing methods in the literature [18], [39], [48], [58] as the authors do not provide the raw dataset or the source code, only the extracted topics and labels. The experimental results show that our method outperforms or provides similar results with both NETL's supervised and unsupervised approaches.

To summarize, in this paper, we propose:

i) a new unsupervised topic labeling method that incorporates the syntactic context dimension (i.e., through the use of Part of Speech filters) into topic labeling, that presents promising results on several datasets;
ii) a new method to process documents for the topic modeling and topic labeling tasks that improve their accuracy;
iii) the use of two new datasets for evaluating the tasks of topic labeling;
iv) the use of three automatic evaluation methods that remove the human annotators from the task of evaluation without affecting the quality of the evaluation.

This paper is structured as follows: Section II presents other research directions related to automatic topic labeling. Section III presents the method as well as the evaluation methodology employed. Section IV presents the datasets as well as the implementation of the proposed method. Section V describes the experimental results. Section VI presents the conclusions and outlines the directions for future work.

II. RELATED WORK

In this section, we discuss several methods that address automatic topic labeling, a research field that has seen a lot of attention in recent years. The aim of topic labeling is to find a relevant label or title for the documents labeled with the same topic [55] and to create a better description that underlines the homogeneity of a given topic [12].

To better understand a topic label, first we will define a topic model z in a text collection D as a probability distribution $p(w|z)$ of words $w \in V$, where V is a vocabulary set [39], among which we can extract the top-k terms, while every document $d \in D$ is associated to a distribution $p(z|d)$ over topics [55]. Thus, we can infer that $\sum_{w \in V} p(w|z) = 1$. Therefore, a topic label or just "label" l for a topic model z is a sequence of semantically meaningful terms that covers the latent meaning of z.

In [39], the authors propose a method for labeling topics that considers this task as an optimization problem. It involves minimizing Kullback-Leibler divergence between word distributions and maximizing the mutual information between a label and a topic model. The evaluation of the method uses two datasets, one containing scientific articles and the other news articles. The methodology employed for evaluation uses automatic as well as human evaluation. The automatic evaluation uses Pointwise Mutual Information, which we also use as our evaluation method. The results show that the proposed labeling methods are quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models.

Similarity measures and topic labeling rules which are specifically designed to find the most agreed labels between

the given topic and the hierarchy are also used for topic labeling [34].

In [59], the authors investigate the impact of using phrase-based topic models for automatic topic labeling. The authors propose a solution based on phrase neural embeddings. Thus, they use a phrase-based topic inference that associates the same topic to each word in a phrase. This inference process is very similar to LDA but with constraints on phrases' topics. Although their experiments show that their solution (both the supervised and the unsupervised version) is efficient and effective compared with a unigram-based topic model, it would have been interesting to have a comparison with other topic extraction models, e.g., LDA or NMF.

In [25], the authors proposed to generate candidate label sets using the top-ranking topic terms from Wikipedia articles. The candidate labels are ranked using a combination of association measures and lexical features. This method is based on the assumption that an appropriate label for a topic can be found among the high-ranking (high probability) terms in that topic [26].

Two effective algorithms that automatically assign concise labels to each topic in a hierarchy by exploiting sibling and parent-child relations among topics were proposed in [37]. The dataset used was collected from the web. The label ranking is done using Term Weighting Based Ranking and the Jensen-Shannon Divergence. The experimental results show that the inter-topic relation is effective in boosting topic labeling accuracy and the proposed algorithms can generate meaningful topic labels that are useful for interpreting the hierarchical topics.

Recently, ontology-based solutions were proposed to solve the task of automatic topic labeling [2], [21]. A method that uses ontologies for automatic topic labeling was proposed in [2]. The authors argue that the quality of topic labeling can be improved by considering ontology concepts rather than words alone. They proposed: i) a topic model that integrates ontological concepts, and ii) a topic labeling method based on the ontological meaning of the concepts included in the discovered topics. The experimental results show that by introducing ontological concepts the generated labels are meaningful.

In [21], the authors go a step further by combining ontology-based labeling and social network analysis. Their proposed ontology-based framework has the following main processing steps: i) LDA topic modeling is performed on the datasets; ii) the resulting topics are used in the social network analysis, i.e., the topics containing the same keywords are connected with a weight value, iii) the influential topics are selected and interpreted based on the proposed ontology. Although it's an interesting solution, the authors do not compare it with other solutions, and they do not say if or how they evaluate its efficiency. To find a label and introduce some supervision, some methods employ external resources such as Wikipedia [5].

For micro-blogging such as Twitter, text summarization using Text Rank [4] as well as hashtags [38] was also used successfully for topic labeling. Other techniques use word vectors and letter trigram vectors to automatically extract labels for topics [22] as well as unsupervised graph-based methods [1], [19], [44]. In the literature, there have also been proposed hybrid methods that make use of multiple measures to increase the chance to find the correct label for the topic [18].

Only a relatively small number of studies have used transformers-based semantic embeddings, e.g., BERT (Bidirectional Encoder Representations from Transformers) [13] or ELMO [43], in the topic analysis.

The impact of using BERT embeddings in topic analysis is tackled in [61] via a model that examines topic evolution in mono and multilingual publications, i.e., English and Chinese. To explore the evolution of different topics in scientific publications, the authors applied LDA to obtain the topic probability value and BERT embedding to analyze topic similarity.

Some of the methods presented in the literature use human evaluation for the quality of a label for a specific topic [2], [5], [25], [26]. The human evaluators are presented with 10 pairings of topic and the candidate label and then asked to rate the label on an ordinal scale of 0-3 where 0 indicates a completely inappropriate label, and 3 indicates a very good label for the given topic [5], [25], [26]. Although this approach is a plausible method for evaluation small corpora and getting a golden standard, this evaluation method is sometimes costly and adds bias to the evaluation process because it is dependent on the evaluator's knowledge and background. Moreover, He *et al.* [62] observed that gold standard labels from human beings suffer from inconsistency. Furthermore, some authors preselect the topics and their labels before presenting them to the human evaluators to improve the quality of the results [2], [5], [25], [26]. We evaluate our proposed model, both automatically and using human annotators. We use the Adjusted Rand Index method for the human evaluation approach. We also automatically label and evaluate the topics as they are found by the proposed unsupervised algorithm. Thus, the quality of the topic labels is estimated automatically using 3 different metrics.

In conclusion, none of the current methods presented in the literature propose a topic labeling technique that also incorporates the syntactic context dimension of the documents belonging to a topic. We address this problem by using a linguistic filter to determine the domain-specific candidate terms for the set of documents belonging to a topic. From the set of candidate terms, the labels are automatically extracted. Moreover, because the candidate terms are the representative terms for the topic as well as for the set of documents belonging to that topic, a label is also a representative term for both topic and set of documents.

III. METHODOLOGY

In this section, we describe the proposed topic labeling method. Our method, TLATR, uses automatic domain-specific term recognition to extract the candidate n-grams

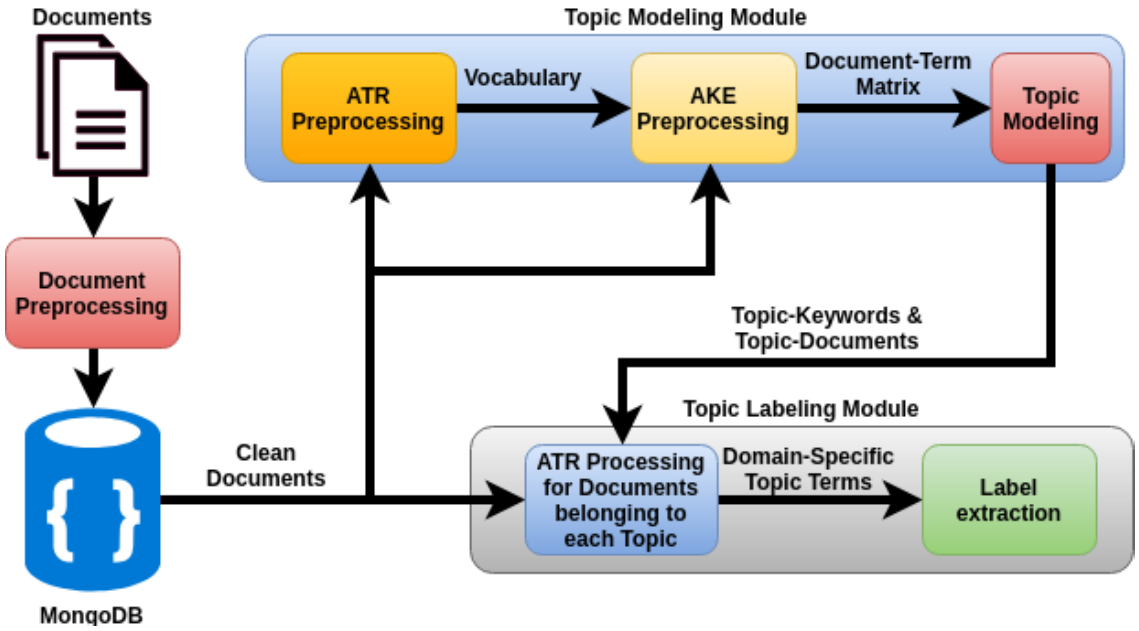

FIGURE 1. Logical schema for the proposed method.

(*cng*) that are relevant to the documents D belonging to a topic z. The candidate n-grams are then scored to obtain the list of relevant domain-specific terms (*dst*). The term with the highest score is chosen from domain-specific terms (*dst*) list as label l for a topic z.

The label's relevance to the topic is then evaluated using Pointwise Mutual Information (PMI), Normalized Pointwise Mutual Information (PMI), and cosine similarity between the vector embedding of the topic's label and its keywords. We also use human evaluators to compute the Average Rating score.

TLATR employs two topic modeling algorithms, i.e., Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). Automatic Keywords Extraction (AKE) methods, i.e., *TF-IDF* and *Okapi BM25*, are used to weight the terms before constructing the document-term matrix used by these algorithms. These two algorithms together with the *TF-IDF* and *Okapi BM25* weighting schemes showed good accuracy using empirical evaluation [52], [54]. Our resulted topic models are then evaluated by applying the Adjusted Rand Index (ARI).

In the following subsections, we briefly describe each module (Figure 1), although we will put more emphasis on the topic modeling algorithm.

A. TOPIC MODELING MODULE

The operation of this module is done in three stages:

(i) *ATR Preprocessing*: during this stage we extract the list of domain-specific terms (*dst*) using Automatic Term Recognition (ATR) and construct the vocabulary;

(ii) *AKE Preprocessing*: this second stage implies building the document-term matrix employing Automatic Keyword Extraction (AKE);

(iii) *Topic Extraction*: in the last stage we apply topic modeling algorithms to extract topics.

1) ATR PREPROCESSING

To construct and minimize the vocabulary for topic modeling, we use ATR. ATR determines specific terms and n-grams relevant to a domain, thus removing all the other words that do not add any information gain. One method used to extract domain-specific multi-word terms is C-Value introduced for the first time in [15]. Equation (1) presents the formula for C-Value, where a is a candidate string, $|a|$ is the number of terms in a, $f(a)$ is the frequency of a in the corpus, T_a is the list of candidate terms (n-grams) that contain a and $P(T_a)$ is the number of these candidate terms (n-grams) [17].

$$\text{C-value(a)} = \begin{cases} \log_2 |a| \cdot f(a) \text{ where a is not nested} \\ \log_2 |a| \cdot (f(a) - \frac{1}{P(T_a)} \\ \cdot \sum_{b \in T_a} f(b)) \text{ otherwise} \end{cases} \quad (1)$$

The linguistic filters used for extracting the candidate string are user-defined and usually, they contain Nouns, Adjectives, and Noun Prepositions [52]. The use of the linguistic filter adds context to the topic labeling process. Originally, C-Value was defined to extract n-grams, with $n \geq 2$, but it can be adapted to also extract single word domain-specific terms by changing the $\log_2 |a|$ into $\log_2(1 + |a|)$ [32]. In this stage, the adapted form to extract the domain-specific terms (*dst*) is implemented to construct the vocabulary, which is afterward used in the second stage for the document-term matrix construction.

Other methods based on C-Value were proposed, e.g., NC-Value that better combines linguistic patterns with the textual context [16], [17]. Furthermore, there are several methods implemented that combine ATR and AKE weights such as F-TFIDF, F-Okapi [32] or LIDF-value [33]. These methods are beyond the scope of this paper.

2) AKE PREPROCESSING

Information Retrieval uses AKE to query textual datasets and return the most relevant documents for a given query [32]. Although keywords are weighted, the same as in ATR, AKE does not employ linguistic patterns and context information for the evaluation of candidate terms (n-grams) to detect domain-specific multi-words [15]. The most relevant use cases for AKE are weighting terms for document classification and clusterization tasks.

For constructing the document-term matrix after the corpus vocabulary is determined, we use *TF-IDF* and *Okapi BM25*. The *TF-IDF* (Equation (2)) is a numerical statistical function intended to reflect how important is a word to a document in a collection or corpus [28]. The *TF-IDF* is computed by multiplying the term frequency *TF* by the inverted document frequency *IDF*.

$$TFIDF(t, d, D) = TF(t, d) \cdot IDF(t, D) \tag{2}$$

Several computational methods for *TF* and *IDF* have been proposed in the literature [41]. To prevent a bias towards long documents, the augmented term frequency (*TF*) was introduced in [47] (Equation 3).

$$TF(t, d) = K + (1 - K) \cdot \frac{f_{t,d}}{\max_{t' \in d}(f_{t',d})} \tag{3}$$

The augmented *TF* uses the number of co-occurrences $f_{t,d}$ of a word in a document, normalized with the frequency of the most frequent term t', i.e., $\max_{t' \in d}(f_{t',d})$. The default value of the free parameter K is usually chosen as 0.5, in absence of advanced optimization [41].

The *IDF* (Equation (4)) measures how much information a term provides for a given corpus [45]. Its main purpose is to determine whether a term is rare or common across all documents. The *IDF* for a term is computed as the logarithm of the fraction obtained by dividing the number of documents in the corpus N by the number of documents containing that term $n = |\{d \in D : t \in d\}|$.

$$IDF(t, D) = 1 + \log \frac{N}{n} \tag{4}$$

The *Okapi BM25* is a bag-of-words retrieval function initially used for obtaining the most relevant documents for a given query in textual datasets [46]. It can also be used to rank the terms in a document corpus and then construct the document-term matrix, used by Text Mining algorithms to extract textual knowledge [60]. *Okapi BM25* is presented in Equation (5), where $||d||$ is document d's length, i.e., the number of terms appearing in d. The average document length for the corpus, $avgDL = avg_{d' \in D}(||d'||)$, is used to remove any bias towards long documents. In absence of advanced optimization, the default values of the free parameters are usually chosen as $k_1 \in [1.2, 2.0]$ and $b = 0.75$ [35], [49], [50].

$$OkapiBM25(t, d, D) = \frac{TFIDF(t, d, D) \cdot (k_1 + 1)}{TF(t, d) + k_1 \cdot (1 - b + b \cdot \frac{||d||}{avgDL})} \tag{5}$$

3) TOPIC EXTRACTION

To extract the topics from the documents' corpus, we use the LDA and NMF algorithms.

LDA [6] is a generative statistical model that groups together terms that are syntactically different but have similar meanings and represent the same concepts. The documents are represented as random mixtures over latent topics. Each topic is characterized by a distribution over words. LDA uses a corpus of M documents $D = \{d_1, d_2, \ldots, d_M\}$. Each document d_i ($i = \overline{1, M}$) is seen as a sequence of N words w_j ($j = \overline{1, N}$), thus $d_i = \{w_1, w_2, \ldots, w_N\}$. LDA models document lengths as Poisson distributions, i.e., $N \sim Poisson(\xi)$. Then computes the probability that a given word w_j in the document d_i belongs to topic z is θ_t ($t = \overline{1, k}$), where $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ is a random multinomial distribution generated by the Dirichlet distribution, i.e., $\theta \sim Dirichlet(\alpha)$, where α is a k-vector with components $\alpha_i > 0$. For each word w_j from the vocabulary, we have the following two steps. First, a topic is chosen as $z_j \sim Multinomial(\theta)$. Second, a word w_j from a multinomial probability conditioned on the topic z_j, i.e., $p(w_j|z_j, \beta)$, is chosen, where β is a $k \times ||V||$ matrix with $||V||$ the dimension of the vocabulary.

NMF is a dimension reduction and factor analysis method used in clustering and topic modeling [23]. Given a non-negative matrix $A \in \mathbb{R}^{m \times n}$, the goal of NMF is to find two matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ having only non-negative entries when desired a lower dimension k. When used for topic modeling, matrix A is the document-term matrix and k represents the number of topics desired, while W is the document-topic matrix and H is the topic-term matrix. The matrices W and H are computed from an optimization problem defined either with the Frobenius norm or the Kullback-Leibler divergence [27], [31] or other divergences [14], [31]. For our implementation, we chose the Frobenius norm as it is widely used in the literature.

B. TOPIC LABELING MODULE

This module consists of two components. The first component applies ATR on each group of documents belonging to each topic, similar to the one discussed in Subsection III-A1, with the difference that here we use the original version of C-Value where only n-grams with $n \geq 2$ are extracted. The second component uses the output of the ATR preprocessing component to extract and select from the domain-specific list the term with the highest score as topic label. Thus, ATR is used for automatic label detection by extracting n-grams using linguistic patterns and the topic

Algorithm 1 TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition

Require: a topic-documents set $TD = \{(z_i, D_i)|D_i \textit{ is the set of documents for topic } z_i\}$ and the list of linguistic patterns LP
Ensure: a topic-label set $TL = \{(z_i, l_i)|l_i \textit{ is the label for topic } z_i\}$

1: $TL = \emptyset$
2: **for each** $(z_i, D_i) \in TD$ **do in parallel**
3: $cng = \emptyset$
4: $dst = \emptyset$
5: **for** $d \in D_i$ **do**
6: $d = expandContractions(d)$
7: **for** $sentence \in d$ **do**
8: $lemmaPOS = \emptyset$
9: $tokens = wordTokenize(sentence)$
10: $tokensPOS = posTokenize(tokens)$
11: **for** $token, pos \in tokensPOS$ **do**
12: $lemma = lemmatization(token, pos)$
13: $lemmaPOS = lemmaPOS \cup \{(lemma, pos)\}$
14: $cng = cng \cup extractNGrams(lemmasPOS, LP)$
15: **for** ngram $\in$ cng **do in parallel**
16: $dst = dst \cup \{(ngram,$ C-Value$(ngram))\}$
17: $dst = sort(dst, key$=C-Value$)$
18: $l_i = dst[0]$
19: $TL = TL \cup \{(z_i, l_i)\}$
20: **return** TL

labels are determined from the list of the domain-specific terms extracted from the documents belonging to the same topic.

The proposed method is implemented through Algorithm 1. The input is a topic-documents set (*TD*) that contains the list of documents assigned to each topic and a list of linguistic patterns (*LP*) needed to extract the candidate n-grams (*cng*) before applying *C-Value* and determining the domain-specific terms (*dst*). The output is a topic-label set (*TL*) that contains the label l_i for a topic z_i. The linguistic patterns (*LP*) has a dual-purpose:

i) adding context to the topic labeling through the use of the part of speech patterns, and
ii) filtering the words taken into account for determining the domain-specific terms (*dst*)

During the iteration of the topic-documents set (*TD*), the candidate n-grams (*cng*) for each topic are extracted. The candidate n-grams (*cng*) list contains all the candidate n-grams for all the documents labeled with the same topic. To extract the candidate n-grams (*cng*), each document assigned to a topic is processed as follows (Lines 5 to 14):

i) the contractions are expanded (Line 6);
ii) the document is split into sentences (Line 7);
iii) for each sentence the tokens and their corresponding part of speech (*pos*) and lemmas are extracted (Lines 8 to 13);
iv) the candidate n-grams (*cng*) is constructed using the list of linguistic patterns (*LP*) and the lemma and part of speech of each term (Line 14).

During the iteration of the candidate n-grams (*cng*) list, C-Value is used to extract and score the domain-specific terms (*dst*) (Lines 15 and 16). The domain-specific terms (*dst*) are then sorted in descending order by their C-Value score (Line 17).

The topic label is determined (Line 18) and the topic-label set (*TL*) is updated (Line 19). When labels are determined for all the topics, the algorithm finishes and returns the topic-label set (*TL*) (Line 20).

C. EVALUATION METHODOLOGY

1) TOPIC MODELING EVALUATION METHODS

For evaluating the performance of the proposed topic modeling module, the Adjusted Rand Index (ARI) cluster evaluation method is employed. ARI is the improved version of the Rand Index (RI), which has the disadvantage that for two random partitions the RI expected value is not a constant. ARI (Definition 1) calculates the fraction of correctly classified and respectively misclassified elements by assuming a generalized hypergeometric distribution as null hypothesis [57].

Definition 1 (Adjusted Rand Index (ARI)): Given a set D of $n = ||D||$ elements and two clustering partitions $X = \{X_1, X_2, \ldots, X_r\}$ $(r = ||x||)$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$ $(s = ||Y||)$, the overlap between X and Y $(X \cap Y)$, can be summarized in a contingency table $[n_{ij}]$, where $n_{ij} = |X_i \cap Y_j|$, $a_i = \sum_{j=1}^{s} n_{ij}$ and $b_j = \sum_{i=1}^{r} n_{ij}$. Equation (6) defines the Adjusted Rand Index using the contingency table.

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \cdot \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right) - \frac{\sum_i \binom{a_i}{2} \cdot \sum_j \binom{b_j}{2}}{\binom{n}{2}}} \tag{6}$$

2) AUTOMATIC TOPIC LABELING EVALUATION METHODS

As evaluation methods for the Topic labeling module, we utilize Pointwise Mutual Information (PMI), Normalized Pointwise Mutual Information (NPMI), and cosine similarity between the average word vectors of a topic label and of its keywords.

PMI determines if a label is relevant to a topic. In computational linguistics, PMI has been used for finding collocations and associations between words [36]. A collocation is a sequence of words or terms that co-occur more often than would be expected by chance. A collocation with a high PMI denotes that the probability of co-occurrence of the two terms is only slightly lower than the probabilities of occurrence of each word. Moreover, PMI has been used for cluster labeling [9], [35].

In the context of topic labeling, PMI (Definition 2) determines the relevance of a label to a topic, treating each label as a collocation. By computing the PMI for a label using all the collocations that appear in the documents belonging to a topic, it can be determined if a label is indeed representative for a topic.

Definition 2 (Pointwise Mutual Information (PMI)): The PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence. This can be described by $pmi(x, y) = log(\frac{p(x,y)}{p(x)\cdot p(y)})$.

The main drawback of PMI is its bounds, which are defined as follows: $-\infty \leq pmi(x, y) \leq min[-log(p(x)), -log(p(y))]$. To solve this problem and normalize the bounds to $[-1, 1]$, the NPMI (Definition 3) was introduced [7]. The interpretation of the NPMI values is:

i) -1 means that the pair of terms never occurring together;
ii) 0 means that the pair of terms are independent;
iii) 1 means that there is a complete co-occurrence for the pair of terms, i.e., they always appear together.

Definition 3 (Normalized Pointwise Mutual Information (NPMI)): Given a pair of outcomes x and y belonging to discrete random variables X and Y, when these two outcomes (x and y) only occur together, the chance of seeing one equals the chance of seeing the other, which equals the chance of seeing them together, meaning that the self-information $h(x, y)$ is $h(x, y) = -log(p(x)) = -log(p(y)) = -log(p(x, y))$. Then the NPMI is defined as $npmi(x, y) = \frac{pmi(x,y)}{h(x,y)}$.

In the context of topic labeling, NPMI normalizes the score given by PMI for a label. So the closest the score is to the upper bound, the more relevant is that label for the set of documents belonging to that topic.

We use word vectors to estimate the semantic similarity between the labels and the topics [20]. Both the topic labels and the keywords for the topic are transformed into word vectors using word embedding, representing each topic label (TL) and topic keywords (TK) as an average of their word vectors [40]. Thus, each term $l_i \in TL$ is transformed into a vector v_{l_i} and then the average vector is computed as $v_l = \frac{1}{n} \cdot \sum_{i \in n} v_{l_i}$, with $i = \overline{1, n}$, $n = ||TL||$. The same steps are applied on the topic keywords, thus each keyword $k_j \in TK$ is transformed into a vector v_{k_i} and the average vector is computed as $v_k = \frac{1}{m} \cdot \sum_{i \in m} v_{k_i}$, with $i = \overline{1, m}$, $m = ||TK||$. Finally, the cosine similarity (Definition 4) is used on the average word vectors (i.e., v_l and v_k) to estimate a similarity score [29], [30].

Definition 4 (Cosine similarity ($cos(\theta)$)): Given two p-dimensional vectors x and y, the similarity between them can be computed using the cosine similarity as follows: $cos(\theta) = \frac{\sum_{i=1}^{p} x_i \cdot i}{\sqrt{\sum_{i=1}^{p} x_i^2} \cdot \sqrt{\sum_{i=1}^{p} y_i^2}}$

3) HUMAN EVALUATION METHODS FOR TOPIC LABELS

To evaluate the quality of TLATR, we compute the Average Rating (AvgR) score by asking human annotator to rate a topic on a 4-point Likert Scale as follows:

i) A perfect label that describes the topic is graded with 3 (Very good label);
ii) A label that is related to the topic but does not completely capture the topic is graded with 2 (Reasonable label);
iii) A label that is semantically related to the topic but does not describe the topic very well is graded with 1 (Semantically related); but would not make a good topic label;
iv) A label that is unrelated to the topic is graded with 0 (Inappropriate label).

Equation (7) presents the formula for computing the Average Rating, where n is the number of annotator and $s_j(z_i, l_i)$ ($j = \overline{1, n}$) is the grade given by each annotator of how well the label l_i describes topic z_i.

$$AvgR(z_i, l_i) = \frac{\sum_{j=1}^{n} s_j(z_i, l_i)}{n} \tag{7}$$

IV. EXPERIMENTAL SETUP

This section is split into three parts. First, we analyse the dataset used for testing. Second, we present the label extraction process. In the last part, we discuss the implementation of TLATR.

A. CORPORA

Two datasets are used for testing the proposed method for automatic topic labeling. Both sets contain scientific articles abstracts. The first set of scientific articles abstracts (ART) was presented in [51] and is publicly available online.[1] The second dataset was collected from Arxiv[2] (ARX) [52], [54]. Both datasets are labeled. The labels are used for setting the number of topics for the topic modeling algorithms. Table 1 presents the number of documents and the number of labels for each corpus.

[1]Scientific articles dataset https://aminer.org/collaboration
[2]Arxiv database https://arxiv.org/

TABLE 1. Corpora.

Corpus	No. Documents	No. Labels
ART	18 464	5
ARX	166 512	19

TABLE 2. Corpora labels.

Corpus	Labels	No. Documents
ART	database	2 270
	data mining	5 059
	medical	3 066
	theory	3 995
	visualization	4074
ARX	cs	21 999
	math	24 000
	physics	12 000
	physics_astro-ph	12 000
	physics_cond-mat	12 000
	physics_gr-qc	7 685
	physics_hep-ex	4 147
	physics_hep-lat	1 883
	physics_hep-ph	9 736
	physics_hep-th	9 572
	physics_math-ph	7 113
	physics_nlin	3 303
	physics_nucl-ex	2 045
	physics_nucl-th	4 255
	physics_physics	12 000
	physics_quant-ph	9 908
	q-bio	3 608
	q-fin	1 399
	stat	7 859

As depicted in Table 2, the classes from the ART corpora are almost disjoint, while the labels from the ARX corpora are not totally disjoint, e.g., the physics labels. Moreover, to determine if the proposed method for automatic topic labeling using automatic term recognition really determines representative labels for each topic, the ARX corpus was deliberately collected to contain labels that are not disjoint.

B. LABEL EXTRACTION

To extract a topic label, the following steps are applied on all the documents labeled with that topic:

i) expand contractions;
ii) extract sentences;
iii) apply part of speech tagging and extract lemmas;
iv) construct the list of candidate n-grams (*cng*) using lemmas;
v) apply C-Value on each candidate n-gram;
vi) set as label for the topic the n-gram with the highest C-Value score.

As discussed previously, the candidate n-grams list contains all the candidate n-grams for all the documents labeled with the same topic.

During the extraction of the candidate terms (n-grams), a linguistic filter is used to extract from 2-grams up to 5-grams for each set of documents. This linguistic filter is defined by the following regular expression $((Noun|Adjective)(Preposition)^{?})^{+}Noun$, and in our algorithm, it is used as the list of linguistic patterns (*LP*). As mentioned before, the linguistic pattern has dual-functionalities: to add context to the proposed topic labeling method and to filter the terms.

C. IMPLEMENTATION

The textual datasets are processed and stored in a MongoDB database. The architecture used for processing and storing documents is presented in detail in [53].

For extracting the topics, we apply the method described in Section III-A. This method extracts topics using contextual cues, adding a context dimension to the topic extraction process.

During text preprocessing phase, we use the *C-Value Vocabulary* (CVV) together with the *TF-IDF* and *Okapi BM25* term weighting functions to create the document-term matrix. The *CVV* employs on each document in the corpus the following steps [52]:

i) expanding the contractions;
ii) extracting the sentences;
iii) applying the part of speech tagging and extract lemmas;
iv) applying C-Value and extracting the candidate terms;
v) removing stop words and punctuation;
vi) removing duplicate terms;
vii) constructing vocabulary.

Table 3 presents the vocabulary size for both datasets. For comparison reasons, we added the *Clean Text Vocabulary* (*CTV*) size, which contains only words (the punctuation characters and stop words were removed).

TABLE 3. Vocabulary size.

Corpus	CTV size	CVV size
ART	33 766	22 666
ARX	112 834	81 997

This vectorization is used because the cluster evaluation scores presented in the results section of this paper are consistent with the ones presented in [54]. Moreover, this vectorization also improves human readability and topic coherence and separability.

We evaluate our solution with the NMF and LDA topic modeling algorithms. We use the Scikit-learn [42] implementation of the topic modeling algorithms.[3] For evaluation, we implemented the PMI and NPMI methods, while for the document similarity we employ the implementation from the Python package SpaCy,[4] that has a full English language model for word embedding.

To compute PMI and NPMI, we apply the following steps:

i) determine the documents that belong to each topic (a document is assigned to a single topic);
ii) extract the n-grams from the subset of documents belonging to a topic;
iii) compute the score using the label and the n-grams.

[3]Scikit-learn `https://scikit-learn.org/stable/`
[4]SpaCy `https://spacy.io/`

TABLE 4. ART NMF TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data mining	data mining method set approach algorithm paper pattern classification large	4.75	0.55	0.80	3.00
2	database system	query database relational performance xml object processing optimization language data	4.89	0.54	0.80	3.00
3	real time	image surface method object model motion point shape scene camera	7.37	0.71	0.65	2.43
4	low bound	problem algorithm time bound graph approximation log number case polynomial	7.29	0.81	0.55	2.65
5	research paper	system information application research technology user management paper data visualization	5.74	0.57	0.87	2.88

TABLE 5. Compassion TLART vs NETL on ART NMF TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data mining	data set	methodology	**0.80**	0.78	0.65
2	database system	relational database	relational database	0.80	**0.84**	**0.84**
3	real time	camera	camera	**0.65**	**0.65**	**0.65**
4	low bound	polynomial	polynomial	0.55	**0.66**	**0.66**
5	research paper	information management	information management	**0.87**	0.84	0.84

To compute the Average Rating (AvgR), we used 30 graduate and undergraduate students of Computer Science. The annotators received the topic label, the topic's top-10 relevant keywords, and a grading scale from 0 to 3. They were asked to grade, using this scale, how well the label describes the corresponding topic.

We compare the topic labels obtained by our method with the labels obtained by NETL [5]. NETL offers two distinct topic labeling models: i) a supervised NETL (SNETL) model, and ii) an unsupervised NETL (UNETL) model. We used the cosine similarity for this comparison, as the task would be too tenuous for human annotators.

The implementation, datasets, and the experimental results are available on Github.[5]

V. EXPERIMENTAL RESULTS

This section presents and analyzes the experimental results of TLATR. We build four sets of tests for each corpus using the *CVV* vectorization with different term weighting schemes as follows:

i) NMF with *TF-IDF*;
ii) LDA with *TF-IDF*;
iii) NMF with *Okapi BM25*;
iv) LDA with *Okapi BM25*.

Previous works have proven that the combination between topic model algorithm and weighting schema is really dependent on the document dataset [52], [53]. The quality of topic modeling can be influenced by the length of documents in the corpus, the frequency of rare words, etc. Thus, we use *TF-IDF* to reward term frequency and penalizes document frequency and we use *Okapi BM25* to take into account the document length and term frequency saturation.

All of the datasets' results are presented in the following subsections.

[5]Github source code: https://github.com/cipriantruica/TL_ATR

A. THE ART CORPUS

Table 4 presents the results obtained when using the NMF topic algorithm with the *TF-IDF* term weighting scheme. All the NPMI values for this setup are over 0.55. As observed, relevant labels are assigned to the two topics that contain terms related to "database systems" and "data mining". The topic with the highest PMI contains terms related to mathematical functions and it has been labeled with "low bound" also a property of mathematical functions. The cosine similarity between the topic's label and its keywords is over 0.55, proving that the method extracts relevant labels for each topic. The highest similarities ($cos(\theta) \geq 0.80$) are obtained for the topics labeled with "data mining", "database systems" and "research paper" while the lowest ($cos(\theta) \leq 65$) for the topics labeled with "real time" and "low bound". The ARI for this test is $\sim$ 0.39 which is consistent with the results in previous works [56]. The first two topics were scored by annotators with a perfect AvgR (3.00). Using this configuration, the AvgR score is over 2.43, obtained for the label that has also the lowest PMI, NPMI, and cosine similarity scores. The high scores denote that both automatic and human evaluation find the labels appropriate for the documents belonging to the extracted topics.

The results of the comparison between our proposed method, TLATR, and NETL, both supervised (SNETL) and unsupervised (UNETL), are listed in Table 5. In the majority of cases, the similarities between these three methods are very close. One exception is with topic 1, where TLATR and SNETL provide better results than UNETL, 0.80 and 0.78 respectively, as opposed to 0.65. The lowest cosine similarities are obtained with all the methods for topics 3 and 4. The high scores obtained by cosine similarity denote that our method manages to accurately label the extracted topics.

Table 6 presents the results obtained when using the LDA topic algorithm with the *TF-IDF* term weighting scheme. With this second configuration, four topics were labeled the same as in the previous case. Another topic related to "graph theory" and "algorithm analysis" was discovered and has been labeled with "approximation algorithm". For this test,

TABLE 6. ART LDA TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data mining	system data research database information paper knowledge application technology medical	5.73	0.61	0.80	3.00
2	database system	data query database performance algorithm paper system result approach technique	5.08	0.50	0.85	3.00
3	real time	image method model surface motion visualization object shape reconstruction approach	7.23	0.70	0.63	2.36
4	low bound	bound algorithm problem time complexity polynomial low number random case	7.25	0.84	0.66	2.67
5	approximation algorithm	graph algorithm problem approximation time edge vertex minimum tree log	5.08	0.57	0.87	2.76

TABLE 7. Compassion TLATR vs NETL on LDA TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data mining	information system	information system	0.80	**0.85**	**0.85**
2	database system	database	database	**0.85**	0.69	0.69
3	real time	visualization computer graphics	approximation	**0.63**	0.62	0.40
4	low bound	polynomial	polynomial	**0.66**	0.57	0.57
5	approximation algorithm	minimum spanning tree	optimization problem	**0.87**	0.69	0.66

TABLE 8. ART NMF Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data mining	data mining visualization rule discovery knowledge association pattern approach technique	4.56	0.53	0.77	3.00
2	database system	database system query object management relational language xml processing performance	4.03	0.47	0.83	3.00
3	real time	image model segmentation motion reconstruction method surface registration object shape	7.47	0.70	0.61	2.37
4	low bound	algorithm problem time graph approximation result number bound tree polynomial	7.50	0.81	0.55	2.36
5	research paper	research information medical technology paper health application care clinical patient	5.60	0.64	0.84	2.85

the lowest cosine similarity ($cos(\theta) \geq 0.60$) is higher than in the previous test. The highest cosine similarity is $cos(\theta) = 0.87$. The ARI for this test is ~ 0.30. Thus, topic labeling for this setup worked better than the one used in the first setup, while the opposite is true for topic modeling. As in the previous case, the topics labeled with ''data mining'' and ''database systems'' were scored by annotators with a perfect AvgR (3.00). The lowest AvgR score (2.36) was obtained by the topic labeled with ''real time''. For Topic 3, we observe that the human annotators did not found a direct connection between the topics keywords and the label, while the automatic evaluation using NPMI found a high similarity between the topic and the label. Furthermore, we observe the reverse for Topic 2, where the human evaluation and cosine similarity obtain higher scores than NPMI. We can conclude that these results are directly impacted by the distribution of words found in the documents belonging to a topic.

As shown in Table 7, for the majority of topics, with one exception, TLATR has the highest cosine similarity of the three methods. As it can be observed, the labels found by TLATR are much more specific compared to the other solutions that generalize more, and thus resulting in lower similarity scores. Based on these scores, we can conclude that i) the extracted labels are good candidates for the topics discovered by LDA when *TF-IDF* is used as a term weight, and ii) our method outperforms NETL for this configuration.

Table 8 presents the results obtained when using the NMF topic algorithm with the *Okapi BM25* term weighting scheme. For this configuration, the labels are the same as in the case of NMF with *TF-IDF* vectorization. The highest cosine similarity was obtained for topic number 5 that has the ''research paper'' label, although, it can be observed that the terms for this topic are more related to research papers from the medical domain. This is a direct impact of using a *Okapi BM25* vectorization: the relevance of terms removes the bias towards longer documents, as this function takes into account the length of the document. For this test the lowest cosine similarity between the topic's label and its keywords is $cos(\theta) = 0.55$, while the highest is $cos(\theta) = 0.84$. These similarity results are inline with the previous results, as is also the case for ARI (~ 0.34). The human annotators gave a perfect AvgR score (3.00) to the topics labeled with ''data mining'' and ''database systems''. For this set of experiments, we observe a high correlation between the cosine similarity and the AvgR scores. We can conclude that the automatic evaluation is in line with human evaluation, and the labels are good candidates for topic labels.

The results of our proposed method are compared with the SNETL and UNETL methods and are listed in Table 9. For the first two topics, all the methods obtain very similar (if not the same) labels. On the other hand, for topics 3 and 4, both NETL methods have more specific labels. We observe that for the topics obtained using NMF with *Okapi BM25* as term weight, TLATR manages to extract more generic labels than NETL. Furthermore, as the candidate labels are determined from the set of documents belonging to a topic and not the keywords, our method has a higher correlation to the documents than to the topic keywords.

Table 10 lists the results for the LDA topic algorithm with the *Okapi BM25* term weighting scheme tests. Using this configuration a new topic labeled with ''black box'' appears, while the topic labeled with ''real time'' from the previous experiments disappears. This happens because one document can belong to multiple topics, and as the weighting scheme

TABLE 9. Compassion TLART vs NETL on NMF Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data mining	data mining	information visualization	**0.77**	**0.77**	0.74
2	database system	relational database	relational database	**0.83**	**0.83**	**0.83**
3	real time	image registration	image registration	0.61	**0.71**	**0.71**
4	low bound	polynomial	polynomial	0.55	**0.65**	**0.65**
5	research paper	medical research	clinical research	0.84	**0.91**	0.85

TABLE 10. ART LDA Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data set	data image method algorithm model approach feature object pattern classification	5.25	0.56	0.74	3.00
2	database system	database data system query management application web information object mining	4.38	0.46	0.88	3.00
3	black box	abstract reconstruction tree visualization preliminary dynamic pet interactive computation data	10.35	0.99	0.77	2.54
4	low bound	algorithm problem graph time bound approximation point polynomial number linear	7.50	0.81	0.58	2.57
5	research paper	medical image research system clinical patient paper neural information registration	6.34	0.72	0.81	2.91

TABLE 11. Compassion TLART vs NETL on LDA Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data set	data model	data model	0.74	**0.81**	**0.81**
2	database system	web application	web application	**0.88**	0.80	0.80
3	black box	data visualization	data visualization	0.77	**0.79**	**0.79**
4	low bound	polynomial	polynomial	0.58	**0.69**	**0.69**
5	research paper	clinical research	clinical research	**0.81**	**0.81**	**0.81**

changes, the terms that have higher importance are now taken into account. For this label, the annotators gave an AvgR score of 2.54, the lowest for this configuration. Moreover, for the same topic keywords the topic label changed from "data mining" to "data set". For this label, the annotators again gave a perfect AvgR score 3.00.

Although the terms for the topic remain the same as in the previous cases, we can infer the following:

i) the set of documents clustered to this topic changes significantly due to the weighting scheme used;
ii) the most relevant domain-specific terms for these documents is "data set".

For this last set of tests with the ART corpus, the lowest cosine similarity between the topic's label and its keywords is $cos(\theta) = 0.58$, while the highest is $cos(\theta) = 0.88$. The ARI for this test is ~ 0.27. We observe that, for this set of experiments, there is a high correlation between the cosine similarity score and the human annotators.

In Table 11, we compare the obtained cosine similarity of the three methods when applying LDA with the *Okapi BM25* term weighting scheme. In this scenario, the two variants of the NETL approach, both supervised and unsupervised, obtained the same results. We cannot declare a clear winner for this set of experiments, as in some cases, all methods have the same similarity, i.e., topics 4 and 5, and in other cases either TLATR outperforms NETL or vice versa. We conclude that, in this case, the topic modeling algorithm and weighting scheme impact in a positive manner TLATR as well as NETL.

B. THE ARX CORPUS

The results for the NMF topic algorithm with the *TF-IDF* term weighting scheme test are listed in Table 12. The scores computed using PMI and NPMI are high. In some cases the value of NPMI equals 1, underlining the topic label relevance to the topic and the set of documents belonging to that topic. Even the labels with the lowest NPMI value are representative to the topic they label, e.g., the keywords for the topic number 14 contain the label terms. One can observe that, in this case, two topics have the same label: topic number 3 and topic number 4, both of them having the same subject, "dark matter". When looking at the representative terms for these two topics, it can be determined that the documents belonging to the two topics are describing the same concepts using different approaches (their cosine similarity is $cos(\theta) = 0.87$). The highest cosine similarity between the topic labels and its keywords is $cos(\theta) = 0.91$ while the lowest is $cos(\theta) = 0.38$. The ARI for this test is ~ 0.21. The lowest AvgR is obtained for topic 8 and 18 labeled with "low bound" and "upper bound" respectively. Some of the labels obtained a perfect AvgR score, especially the ones that contained the label's words among the topic's top-10 keywords, e.g., topics 1, 3, 12, etc. These evaluation scores are in line with the observations and conclusions drawn for the ATR corpus when using NMF in combination with *TF-IDF*.

By analysing the results of TLATR and the two variants of the NETL for this set of experiments (Table 13), we have the following observations. Considering only SNETL and UNETL, the supervised variant either has the best results or

TABLE 12. ARX NMF TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	black hole horizon gravitational gravity entropy mass spacetime kerr binary	5.97	1.00	0.85	3.00
2	cross section	collision energy production section proton cross momentum transverse hadron tev	7.86	1.00	0.77	2.76
3	dark matter	higgs mass boson standard neutrino decay lhc model search gev	7.66	0.95	0.67	3.00
4	dark matter	cosmological universe matter dark inflation cosmic scale cosmology background energy	6.77	0.81	0.75	3.00
5	differential equation	equation solution differential nonlinear wave condition method problem function schr	5.78	0.62	0.85	2.96
6	field theory	theory field gauge symmetry gravity scalar action term conformal dimension	4.34	0.54	0.83	2.91
7	form factor	quark qcd meson mass pi chiral lattice decay bar heavy	8.81	0.93	0.77	2.79
8	low bound	algorithm problem graph optimization complexity number optimal time bound method	7.17	0.69	0.56	2.23
9	magnetic field	magnetic spin electron field effect interaction energy surface material electronic	6.13	0.72	0.81	2.93
10	mathbb r	space group algebra class manifold operator mathbb representation theorem case	6.77	0.67	0.48	2.39
11	monte carlo	distribution random method function probability model estimation parameter estimator data	10.12	1.00	0.45	2.31
12	neural network	network neural task image art feature deep learning performance convolutional	6.31	0.76	0.81	3.00
13	numerical simulation	dynamic time system simulation model flow particle equilibrium process dynamical	7.53	0.72	0.76	2.86
14	phase transition	phase transition temperature critical order lattice point diagram finite behavior	5.11	0.65	0.79	2.72
15	quantum system	quantum state system entanglement classical qubit mechanic information measurement entropy	3.26	0.34	0.91	2.97
16	single photon	optical frequency photon laser mode light high wave beam cavity	6.16	0.62	0.73	2.75
17	social network	data analysis research paper approach application recent information development tool	6.23	0.57	0.72	2.31
18	upper bound	channel rate capacity user scheme network transmission communication receiver information	8.23	0.81	0.38	2.23
19	x ray	star galaxy stellar mass observation emission ray formation line gas	7.98	0.89	0.51	2.43

TABLE 13. Compassion TLART vs NETL on ARX NMF TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	gravity	gravity	**0.85**	0.75	0.75
2	cross section	kinetic energy	pair production	**0.77**	0.67	0.47
3	dark matter	standard model	standard model	**0.67**	0.44	0.44
4	dark matter	cosmic microwave background	cosmology	**0.75**	0.71	0.68
5	differential equation	differential equation	differential equation	**0.85**	**0.85**	**0.85**
6	field theory	scalar field	scalar field	**0.83**	0.71	0.71
7	form factor	quark	quark	**0.77**	0.69	0.69
8	low bound	time complexity	time complexity	0.56	**0.77**	**0.77**
9	magnetic field	magnetic field	electron	**0.81**	**0.81**	0.69
10	mathbb r	linear algebra	representation theory	0.48	0.70	**0.72**
11	monte carlo	probability distribution	probability distribution	0.45	**0.81**	**0.81**
12	neural network	artificial neural network	artificial neural network	**0.81**	0.79	0.79
13	numerical simulation	dynamical system	dynamical system	0.76	**0.82**	**0.82**
14	phase transition	phase diagram	phase diagram	**0.79**	0.78	0.78
15	quantum system	quantum entanglement	quantum entanglement	**0.91**	0.78	0.78
16	single photon	photon	photon	**0.73**	0.66	0.66
17	social network	information	information	**0.72**	0.69	0.69
18	upper bound	computer network	telecommunication	0.38	**0.75**	0.54
19	x ray	star formation	star formation	0.51	**0.73**	**0.73**

the same with the unsupervised variant, with one exception, topic 10 where UNETL slightly outperforms SNETL. In general, TLATR outperforms or has the same similarity with SNETL, with some exceptions, where the score of the proposed method is much lower.

Table 14 presents the results obtained when using the LDA topic algorithm with the *TF-IDF* term weighting scheme. For this configuration, all the topic labels have an NPMI value over 0.50. The high values of NPMI denote that the terms belonging to the label also appear frequently within the documents belonging to the topic.

As in the previous test, some topics have the same label, e.g., ''black hole'' (for topics 1 and 2), ''low bound'' (for topics 8 and 9) and ''magnetic field'' (for topics 10 and 11). This is a direct impact of the topic modeling algorithm as it uses a soft clustering approach. Due to this property the set of documents belonging to a topic changes. Thus, the topic labeling method takes into account different candidate terms from which to choose a label.

The highest cosine similarity between the topic's labels and its keywords is 0.82, while the lowest is 0.41. The ARI for this test is ~ 0.21. These scores denote that the context included through the use of word embeddings plays an important role in better labeling the topics.

The lowest AvgR is obtained for labels ''low bound'' (for both topics) and ''constacyclic code''. We observe that some of the labels that obtained a perfect AvgR score in the previous tests and reappeared in this set of tests are also scored by the annotators with an AvgR $= 3.00$. For the labels that obtain AvgR scores over 2.75, we also observe that there is a high correlation between the label and the topic's keywords. This correlation is underlined by the high scores obtained through

TABLE 14. ARX LDA TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	star galaxy ray mass observation emission stellar source high line	9.42	1.00	0.76	3.00
2	black hole	theory field gauge black symmetry hole gravity quantum scalar equation	8.18	1.00	0.81	3.00
3	constacyclic code	phys rev bf lett al textbf paper fr uhlenbeck wavelet	10.41	0.92	0.76	2.21
4	cross section	quark qcd collision heavy calculation nuclear energy momentum nucleon meson	9.01	1.00	0.75	2.79
5	dark matter	dark cosmological universe matter inflation model scale planck cosmology cosmic	6.81	0.81	0.78	3.00
6	differential equation	equation function operator solution matrix problem graph eigenvalue result polynomial	7.18	0.67	0.83	2.93
7	functional theory	molecule molecular hydrogen van calculation ab li atom initio functional	7.89	0.80	0.81	2.71
8	low bound	graph network node model complexity logic population problem algorithm tree	8.24	0.73	0.51	2.21
9	low bound	channel information code capacity communication rate bound scheme bit error	6.82	0.68	0.61	2.27
10	magnetic field	system dynamic equation equilibrium particle numerical phase model time energy	7.34	0.75	0.76	2.71
11	magnetic field	spin magnetic temperature phase topological electron band state transition lattice	6.13	0.67	0.74	3.00
12	moduli space	group algebra manifold space lie class algebraic mathbb invariant representation	7.47	0.69	0.72	2.77
13	molecular dynamic	cell surface protein material elastic molecular concentration force simulation polymer	7.11	0.62	0.75	2.88
14	monte carlo	method algorithm problem estimation distribution data error model approach function	10.56	1.00	0.64	2.67
15	neural network	network task image neural art method data learning feature datasets	6.54	0.74	0.80	3.00
16	quantum mechanic	physic development research recent question theory understanding year mechanic concept	6.89	0.71	0.78	3.00
17	single photon	quantum optical photon frequency laser atom state light system mode	6.22	0.60	0.75	2.83
18	social network	network user paper performance strategy market data price system time	5.71	0.50	0.69	2.79
19	standard model	decay neutrino mass gev higgs lhc tev detector boson standard	6.81	0.79	0.69	2.43

TABLE 15. Compassion TLART vs NETL on ARX LDA TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	star	stellar evolution	**0.76**	0.55	0.68
2	black hole	quantum field theory	quantum field theory	0.81	**0.84**	**0.84**
3	constacyclic code	mathematical physics	mathematical physics	**0.76**	0.31	0.31
4	cross section	nucleon	nucleon	**0.75**	0.66	0.66
5	dark matter	cosmology	cosmology	**0.78**	0.73	0.73
6	differential equation	schrodinger equation	rotation matrix	**0.83**	0.54	0.66
7	functional theory	molecule	molecule	**0.81**	0.75	0.75
8	low bound	time complexity	computational complexity theory	0.51	0.69	**0.71**
9	low bound	bit rate	communications protocol	0.61	**0.72**	0.63
10	magnetic field	partial differential equation	euler equations fluid dynamics	**0.76**	0.75	0.72
11	magnetic field	electron	electron	**0.74**	0.68	0.68
12	moduli space	lie algebra	lie algebra	0.72	**0.77**	**0.77**
13	molecular dynamic	polymer	polymer	**0.75**	0.60	0.60
14	monte carlo	normal distribution	normal distribution	0.64	**0.69**	**0.69**
15	neural network	artificial neural network	database	0.80	**0.81**	0.60
16	quantum mechanic	research	research	**0.78**	0.74	0.74
17	single photon	laser	laser	**0.75**	0.69	0.69
18	social network	market data	market data	0.69	**0.85**	**0.85**
19	standard model	neutrino	neutrino	0.69	**0.71**	**0.71**

TABLE 16. ARX NMF Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	black hole horizon gravitational gravity entropy spacetime ad kerr schwarzschild	5.95	1.00	0.82	3.00
2	correlation function	function operator term formula matrix expansion expression integral form coefficient	5.57	0.54	0.82	2.87
3	cross section	collision production quark energy section proton cross momentum heavy hadron	8.33	1.00	0.78	2.78
4	dark matter	mass higgs standard boson decay neutrino model lhc search gev	7.65	0.95	0.67	3.00
5	dark matter	cosmological universe inflation matter scalar dark gravity scale model energy	6.70	0.76	0.77	3.00
6	differential equation	equation solution differential nonlinear condition schr wave existence odinger problem	5.77	0.63	0.86	3.00
7	field theory	theory field gauge symmetry action gravity conformal effective loop dimension	4.37	0.54	0.83	3.00
8	lie algebra	group algebra representation lie finite algebraic module category product subgroup	5.56	0.57	0.75	2.76
9	low bound	graph number vertex edge bound set degree random tree polynomial	7.43	0.75	0.65	2.23
10	magnetic field	optical frequency photon laser mode light wave high beam cavity	6.98	0.66	0.75	2.84
11	magnetic field	phase transition temperature magnetic spin lattice density interaction state critical	6.45	0.72	0.73	2.87
12	moduli space	space manifold dimensional surface dimension condition geometry theorem point curvature	6.49	0.64	0.74	2.67
13	monte carlo	distribution random probability model process estimator parameter estimation statistical asymptotic	10.23	1.00	0.45	2.63
14	neural network	network neural task data feature art learning image model deep	6.60	0.71	0.80	3.00
15	numerical simulation	dynamic system time particle simulation equilibrium dynamical numerical evolution model	7.51	0.71	0.84	2.97
16	optimization problem	algorithm problem method optimization paper computational approach technique performance numerical	5.29	0.52	0.85	2.95
17	quantum system	quantum state entanglement classical system mechanic qubit information measurement entropy	3.35	0.36	0.91	3.00
18	upper bound	channel rate capacity scheme information communication user transmission receiver code	8.08	0.80	0.38	2.25
19	x ray	star galaxy observation ray mass stellar emission formation source high	7.96	0.87	0.52	2.49

the use of NPMI and cosine similarity, denoting that the label is a good representation of the documents belonging to the topic.

Considering this setup, in Table 15, we compare the cosine similarity of all the analyzed methods. At large, TLATR outperforms both variants of NETL, with several

TABLE 17. Compassion TLART vs NETL on ARX NMF Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	black hole	gravitational field	**0.82**	**0.82**	0.73
2	correlation function	wave function	wave function	**0.82**	0.75	0.75
3	cross section	pair production	pair production	**0.78**	0.54	0.54
4	dark matter	standard model	standard model	**0.67**	0.44	0.44
5	dark matter	dark matter	cosmological constant	**0.77**	**0.77**	0.75
6	differential equation	differential equation	differential equation	**0.86**	**0.86**	**0.86**
7	field theory	quantum field theory	lagrangian field theory	**0.83**	0.79	0.64
8	lie algebra	lie algebra	algebraic group	0.75	0.75	**0.81**
9	low bound	chromatic polynomial	chromatic polynomial	**0.65**	0.57	0.57
10	magnetic field	photon	photon	**0.75**	0.66	0.66
11	magnetic field	magnetic field	magnetization	**0.73**	**0.73**	0.46
12	moduli space	curvature	curvature	**0.74**	0.64	0.64
13	monte carlo	probability distribution	probability distribution	0.45	**0.84**	**0.84**
14	neural network	artificial neural network	artificial neural network	**0.80**	0.75	0.75
15	numerical simulation	physical system	euler equations fluid dynamics	**0.84**	0.69	0.68
16	optimization problem	optimization problem	optimization problem	**0.85**	**0.85**	**0.85**
17	quantum system	quantum entanglement	quantum entanglement	**0.91**	0.86	0.86
18	upper bound	information	information	0.38	**0.63**	**0.63**
19	x ray	star formation	star formation	0.52	**0.73**	**0.73**

TABLE 18. ARX LDA Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	star mass gravitational binary hole black disk radius galaxy rotation	8.07	1.00	0.81	3.00
2	black hole	gravity field scalar black hole cosmological theory einstein spacetime gravitational	7.83	1.00	0.83	3.00
3	cross section	collision quark production heavy decay mass section lhc gev energy	8.71	1.00	0.79	2.78
4	dark matter	dark neutrino matter cosmic background experiment data ray mass energy	7.09	0.92	0.75	3.00
5	differential equation	equation solution system time function numerical dynamic nonlinear differential condition	6.76	0.66	0.88	3.00
6	field theory	theory gauge symmetry loop field fermion model lattice operator chiral	5.07	0.57	0.82	3.00
7	graph g	graph entropy edge vertex entanglement random number tree state node	4.44	0.42	0.58	3.00
8	lie algebra	algebra group manifold space algebraic representation lie invariant construction class	6.11	0.58	0.76	2.87
9	low bound	bound proof function problem polynomial set mathbb result theorem paper	7.53	0.73	0.65	2.24
10	macwilliams identity	math triangle cusp adequate cc wedge cd versa angle vice	10.73	0.98	0.68	2.67
11	magnetic field	magnetic electron spin field temperature optical phase effect state transition	6.34	0.67	0.80	3.00
12	nash equilibrium	optimal strategy problem game market cost paper price time agent	9.49	0.84	0.79	2.77
13	neural network	method algorithm data problem approach performance estimation model paper task	7.88	0.82	0.73	3.00
14	numerical simulation	flow simulation force fluid dynamic cell surface model pressure particle	7.21	0.67	0.71	2.89
15	quantum system	quantum state carlo monte system measurement coherent photon qubit single	3.61	0.37	0.85	3.00
16	real time	development research physic science technology application year progress design challenge	7.04	0.59	0.60	2.67
17	social network	network population human individual social model data pattern activity dynamic	5.29	0.50	0.84	3.00
18	upper bound	channel code capacity communication rate transmission bit scheme error information	8.03	0.80	0.39	2.27
19	x ray	emission observation star line ray solar high source galaxy telescope	8.09	0.90	0.52	2.48

exceptions. As such, for the topics 2, 9, 12, 14, 15, and 18 the similarity of TLATR is slightly lower than the best of the other two methods. For some results, the difference is imperceptible, i.e., 0.80 for TLATR vs 0.81 for SNETL. TLATR is clearly outperformed only for topics 8 and 19.

Table 16 shows the results obtained when using NMF with *Okapi BM25*. As in the previous cases, there are some topics that have the same label. Therefore, there are hidden latent semantic meanings between the documents belonging to these topics, even if there is no intuitive and evident relation between the relevant terms that describe a topic. Moreover, due to the change in weighting scheme (which removes the bias towards the document length) some topics are replaced by other topics deem more relevant by the vectorization process. The highest cosine similarity between the topic's labels and its keywords is 0.85, while the lowest is 0.37. The ARI for this test is $\sim$ 0.17. The highest AvgR score is obtained for the labels that contain at least one of the component words among the topic's top-10 keywords. Moreover, we observe that labels that contain similar words or hint at words contained in the topic's top-10 words obtained a perfect score following the annotation process, e.g., Topic 4. For this set of experiments, we observe that the cosine similarity and the AvgR score are correlated, while the NPMI presents a different picture, e.g., Topic 18, Topic 17, etc. We can conclude that the use of word embeddings when computing cosine similarity adds contextual information. Context is also inferred by human annotators when analyzing the similarity between a topic and a label. NPMI does not take into account context when computing the score, as it only analyzes the co-occurrence of words within the text.

By analyzing the results listed in Table 17, it can be observed that the best scores are obtained by TLATR, with only a few exceptions. Thus, we conclude that TLATR's

TABLE 19. Compassion TLART vs NETL on LDA NMF Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	binary star	binary star	**0.81**	0.70	0.70
2	black hole	gravitational field	gravitational field	**0.83**	0.80	0.80
3	cross section	radioactive decay	radioactive decay	**0.79**	0.66	0.66
4	dark matter	cosmic microwave background	cosmic microwave background	0.75	**0.77**	**0.77**
5	differential equation	differential equation	differential equation	**0.88**	**0.88**	**0.88**
6	field theory	quantum field theory	supersymmetry	**0.82**	0.75	0.46
7	graph g	vertex cover	vertex cover	**0.58**	0.47	0.47
8	lie algebra	lie algebra	lie algebra	**0.76**	**0.76**	**0.76**
9	low bound	polynomial	polynomial	**0.65**	0.61	0.61
10	macwilliams identity	right triangle	right triangle	**0.68**	0.70	0.70
11	magnetic field	magnetic field	electromagnetic field	**0.80**	**0.80**	0.72
12	nash equilibrium	optimization problem	optimization problem	**0.79**	0.66	0.66
13	neural network	mathematical model	optimization problem	0.73	**0.74**	**0.74**
14	numerical simulation	fluid dynamics	fluid dynamics	0.71	**0.82**	**0.82**
15	quantum system	quantum system	quantum entanglement	**0.85**	**0.85**	0.71
16	real time	design research	design research	0.60	**0.84**	**0.84**
17	social network	social network	social network	**0.84**	**0.84**	**0.84**
18	upper bound	channel state information	channel state information	0.39	**0.79**	**0.79**
19	x ray	hubble space telescope	hubble space telescope	0.52	**0.66**	**0.66**

process for selecting candidate terms is an important factor for choosing the right label for a topic.

Table 18 presents the results for LDA with *Okapi BM25*. Using this setup, only two topics have the same label. In this case, some topics were labeled with quite interesting labels, better describing the topics and also extracting hidden knowledge about the documents belonging to these topics, e.g., ''macwilliams identity'' for Topic 10, ''nash equilibrium'' for Topic 12, and ''lie algebra'' for Topic 8.

The highest cosine similarity between the topic label and its keywords is 0.88, while the lowest is 0.35. The ARI for this test is $\sim$ 0.19. These results, as in the previous tests, are consistent with the results obtained in the literature [52], [54], [56]. For this set of experiments, we can infer that the annotators scored similarly as they did for the other tests. Furthermore, we observe that there is a high correlation between the scores obtained through automatic evaluation and the ones obtained by human annotators.

The comparison for the last set of experiments done on the ARX corpus is highlighted in Table 19. As observed by analysing the results, topics 18 and 19 scores the worst with our solution. This trend is maintained for all the experiments done on the ARX corpora, with one exception, i.e., when applying LDA with *TF-IDF* which provided better scores. Otherwise, TLATR outperforms or has almost the same similarity scores with the other two.

Table 20 presents the time performance for each extraction task by dataset in seconds. The performance of TLART is highly dependent on the number of records and the documents' length in each dataset. We observe that the labeling task is computationally intensive. Firstly, the algorithm extracts the subsets of documents belonging to each topic. Secondly, the candidate n-grams are determined and the C-Value is computed. Thirdly, the best label is selected from the candidate n-grams. Lastly, the evaluation metrics PMI and NPMI for the extracted labels are computed using the n-grams that are contained in the subset of documents belonging to the topic. Although in our implementation we use process distribution on a single machine, the time performance can be improved using different distributed technologies, such as Spark or Hadoop MapReduce.

TABLE 20. Time performance for each extraction task by dataset.

Topic Modeling	Weighting Schema	Extraction Task	Runtime ART (s)	Runtime ARX (s)
NMF	TF-IDF	Topics	1.20	116.40
		Candidate N-Grams	30.23	492.24
		Labels	3 063.32	114 136.98
	Okapi BM25	Topics	1.00	117.05
		Candidate N-Grams	31.08	518.82
		Labels	3 704.33	114 682.11
LDA	TF-IDF	Topics	110.00	2818.64
		Candidate N-Grams	31.71	499.61
		Labels	3 032.70	116 656.15
	Okapi BM25	Topics	106.91	2 833.10
		Candidate N-Grams	31.79	543.37
		Labels	3 880.41	112 439.57

VI. CONCLUSION

In this article, we present TLATR a new method for topic labeling using automatic domain-specific term recognition. TLATR adds a new context dimension to topic labeling by using linguistic filtering to extract candidate terms and C-Value to score and extract the domain-specific terms. Furthermore, TLATR uses the corpus of documents to extract labels, without the use of external sources as in some state-of-the-art solutions [2], [5], [25].

Our experimental evaluation shows promising results. All the labels detected with the new method are relevant to the extracted topics, as the PMI and NPMI scores and the cosine similarity between the label and the topic's keywords are high. Some labels (e.g., ''lie algebra'', ''nash equilibrium'', etc.) even manage to discover hidden knowledge and to better describe the topics and the documents belonging to them. The duplicate labels that describe multiple topics manage to detect hidden patterns and semantic similarity between the documents that belong to these different topics.

We also employ 30 graduate and undergraduate students to annotate the quality of the label for each topic to evaluate TLATR. Based on the annotators' scores, we compute the Average Rating. We observe that the annotators scored higher the labels which terms are among the topics' relevant keywords.

From a clustering point of view, topic modeling is a soft clustering multi-label process. This feature, together with the weighting scheme we employed, influences the set of documents belonging to a topic, which has a direct impact on the labeling process. This can be avoided by automatically determining the number of topics to be extracted.

Our proposed method was compared with a state-of-the-art solution, NETL [5], that offers two implementations, supervised (SNETL) and unsupervised (UNETL), respectively. In general, TLATR outperforms or has the same similarity with both SNETL and UNETL. Since TLATR extracts the labels directly from the dataset, it finds more specific labels. On the other hand, NETL uses external sources which usually generalises the labels. Nevertheless, sometimes TLATR is outperformed by NETL. Although TLATR finds relevant labels even in such situations, these labels are dependent on the subset of documents belonging to the topic, thus limiting the list of candidate terms. If the topic modeling algorithm miss-groups the documents, the list of candidate terms contains n-grams that generalize the topic's context.

Regarding the results, we can conclude that both the topic algorithm and the weighting schema make the labeling process more vulnerable. The weighting schema emphasizes some terms over others, changing the document set belonging to each topic. *TF-IDF* rewards term frequency and penalizes document frequency, while *Okapi BM25* takes into account the document length and term frequency saturation. On the other hand, the topic modeling algorithms influence the selection of the weighted term based on their heuristics, i.e., Gibbs Sampling or Matrix Factorization. Thus, the experiments show different results and evaluation scores depending on the weighting schema - topic algorithm pairs, as the documents subsets belonging to a topic, and consequently, the topic labels change.

As future work, we plan to improve the automatic topic labelling method by using other ATR methods for extracting the candidate and the domain-specific terms. To further improve topic labeling, we also intend to fine tune the topic modeling algorithms' hyperparameters with parameter optimization. We also plan to employ word, sentence, and document level transformers, e.g., BERT [13], ALBERT [24], XLNet [63], as embeddings. These transformers can enrich the applied topic model, i.e., LDA or NMF. Thus these topic models can take advantage of the contextual information extracted using the transformers in order to capture the topics' semantics more accurately and improve the topic labeling process. Furthermore, methods that automatically determine the number of topics as described in [3], [8] can be used to better determine the topics and improve TLATR.

REFERENCES

[1] N. Aletras and M. Stevenson, ''Labelling topics using unsupervised graph-based methods,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 631–636.

[2] M. Allahyari and K. Kochut, ''Automatic topic labeling using ontology-based topic models,'' in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 259–264.

[3] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. N. Murthy, ''On finding the natural number of topics with latent Dirichlet allocation: Some observations,'' in *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, 2010, pp. 391–402.

[4] A. E. C. Basave, Y. He, and R. Xu, ''Automatic labelling of topic models learned from Twitter by summarisation,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 618–624.

[5] S. Bhatia, J. H. Lau, and T. Baldwin, ''Automatic labelling of topics with neural embeddings,'' in *Proc. Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 953–963.

[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] G. Bouma, ''Normalized (pointwise) mutual information in collocation extraction,'' in *Proc. German Soc. Comput. Linguistics Lang. Technol.*, 2009, pp. 31–40.

[8] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, ''Metagenes and molecular pattern discovery using matrix factorization,'' *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 12, pp. 4164–4169, Mar. 2004.

[9] D. Carmel, H. Roitman, and N. Zwerdling, ''Enhancing cluster labeling using wikipedia,'' in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2009, pp. 139–146.

[10] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, ''Experimental explorations on short text topic mining between LDA and NMF based schemes,'' *Knowl.-Based Syst.*, vol. 163, pp. 1–13, Jan. 2019.

[11] J. Choo, C. Lee, C. K. Reddy, and H. Park, ''UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013.

[12] M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, and J. Han, ''Automatic construction and ranking of topical keyphrases on collections of short documents,'' in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 398–406.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''Bert: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186.

[14] I. S. Dhillon and S. Sra, ''Generalized nonnegative matrix approximations with Bregman divergences,'' in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 283–290.

[15] K. T. Frantzi and S. Ananiadou, ''Automatic term recognition using contextual cues,'' in *Proc. 3rd DELOS Workshop*, 1997, pp. 18–26.

[16] K. T. Frantzi and S. Ananiadou, ''The C-value/NC-value domain-independent method for multi-word term extraction,'' *J. Natural Lang. Process.*, vol. 6, no. 3, pp. 145–179, 1999.

[17] K. Frantzi, S. Ananiadou, and H. Mima, ''Automatic recognition of multi-word terms: The C-value/NC-value method,'' *Int. J. Digit. Libraries*, vol. 3, no. 2, pp. 115–130, Aug. 2000.

[18] A. Gourru, J. Velcin, M. Roche, C. Gravier, and P. Poncelet, ''United we stand: Using multiple strategies for topic labeling,'' in *Natural Language Processing and Information Systems*. Cham, Switzerland: Springer, 2018, pp. 352–363.

[19] D. He, M. Wang, A. M. Khattak, L. Zhang, and W. Gao, "Automatic labeling of topic models using graph-based ranking," *IEEE Access*, vol. 7, pp. 131593–131608, 2019.

[20] T. Kenter and M. de Rijke, "Short text similarity with word embeddings," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 1411–1420.

[21] H. H. Kim and H. Y. Rhee, "An ontology-based labeling of influential topics using topic network analysis," *J. Inf. Process. Syst.*, vol. 15, no. 5, pp. 1096–1107, 2019.

[22] W. Kou, F. Li, and T. Baldwin, "Automatic labelling of topic models using word vectors and letter trigram vectors," in *Proc. Asia Inf. Retr. Symp.* Cham, Switzerland: Springer, 2015, pp. 253–264.

[23] D. Kuang, J. Choo, and H. Park, "Nonnegative matrix factorization for interactive topic modeling and document clustering," in *Partitional Clustering Algorithms*. Cham, Switzerland: Springer, 2015, ch. 7, pp. 215–243. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-09259-1_7#citeas

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–17.

[25] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2011, pp. 1536–1545.

[26] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best topic word selection for topic labelling," in *Proc. Int. Conf. Comput. Linguistics*, 2010, pp. 605–613.

[27] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2000, pp. 535–541.

[28] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[29] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 302–308.

[30] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 211–225, Dec. 2015.

[31] L. Li, G. Lebanon, and H. Park, "Fast bregman divergence NMF using Taylor expansion and coordinate descent," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 307–315.

[32] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Combining C-value and keyword extraction methods for biomedical terms extraction," in *Proc. Int. Symp. Lang. Biol. Med.*, 2013, pp. 45–49.

[33] J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Yet another ranking function for automatic multiword term extraction," in *Proc. Int. Conf. Natural Lang. Process.* Cham, Switzerland: Springer, 2014, pp. 52–64.

[34] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, "Automatic labeling of topics," in *Proc. 9th Int. Conf. Intell. Syst. Design Appl.*, 2009, pp. 1227–1232.

[35] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[36] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[37] X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, and X. Li, "Automatic labeling hierarchical topics," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 2383–2386.

[38] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 889–892.

[39] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 490–499.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–12.

[41] G. Paltoglou and M. Thelwall, "A study of information retrieval weighting schemes for sentiment analysis," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2010, pp. 1386–1395.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[43] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. NAACL*, 2018, pp. 1–15.

[44] P. Pham, P. Do, and C. D. Ta, "Automatic topic labelling for text document using ontology of graph-based concepts and dependency graph," *Int. J. Bus. Inf. Syst.*, vol. 36, no. 2, pp. 221–253, 2021.

[45] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *J. Documentation*, vol. 60, no. 5, pp. 503–520, Oct. 2004.

[46] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. Text Retr. Conf.*, 1995, pp. 109–126.

[47] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[48] I. Sorodoc, J. H. Lau, N. Aletras, and T. Baldwin, "Multimodal topic labelling," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 701–706.

[49] J. K. Spärck, W. Robertson, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments: Part 1," *Inf. Process. Manage.*, vol. 36, no. 6, pp. 779–808, Nov. 2000.

[50] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments," *Inf. Process. Manage.*, vol. 36, no. 6, pp. 809–840, Nov. 2000.

[51] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 1285–1293.

[52] C.-O. Truică, E. S. Apostol, and C. A. Leordeanu, "Topic modeling using contextual cues," in *Proc. 19th Int. Symp. Symbolic Numeric Algorithms Sci. Comput. (SYNASC)*, Sep. 2017, pp. 203–210.

[53] C.-O. Truică, J. Darmont, and J. Velcine, "A scalable document-based architecture for text analysis," in *Proc. Int. Conf. Adv. Data Mining Appl.* Cham, Switzerland: Springer, 2016, pp. 481–494.

[54] C.-O. Truică, F. Radulescu, and A. Boicea, "Comparing different term weighting schemas for topic modeling," in *Proc. 18th Int. Symp. Symbolic Numeric Algorithms Sci. Comput. (SYNASC)*, Sep. 2016, pp. 307–310.

[55] J. Velcin, A. Gourru, E. Giry-Fouquet, C. Gravier, M. Roche, and P. Poncelet, "Readitopics: Make your topic models readable via labeling and browsing," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 5874–5876.

[56] J. Velcin, M. Roche, and P. Poncelet, "Shallow text clustering does not mean weak topics: How topic identification can leverage bigram features," in *Proc. Data Mining Natural Lang. Process. (DMNLP)*, 2016, pp. 25–32.

[57] S. Wagner and D. Wagner, "Comparing clusterings: An overview," ETH Zurich, Zürich, Switzerland, Tech. Rep., 2007, pp. 1–19. [Online]. Available: https://publikationen.bibliothek.kit.edu/1000011477

[58] X. Wan and T. Wang, "Automatic labeling of topic models using text summaries," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2297–2305.

[59] J. Wang and K. Uchino, "Automatic topic labeling for facilitating interpretability of online learning materials," in *Proc. Int. Conf. Web-Based Learn.* Cham, Switzerland: Springer, 2019, pp. 267–273.

[60] J. S. Whissell and C. L. A. Clarke, "Improving document clustering using okapi BM25 feature weighting," *Inf. Retr.*, vol. 14, no. 5, pp. 466–487, Oct. 2011.

[61] Q. Xie, X. Zhang, Y. Ding, and M. Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *J. Informetrics*, vol. 14, no. 3, Aug. 2020, Art. no. 101055.

[62] J. Yang, W. Xu, and S. Tan, "Task and data designing of sentiment sentence analysis evaluation in COAE2014," *Shanxi Univ., Natural Sci. Edit*, vol. 1, no. 3, pp. 16–23, 2015.

[63] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

CIPRIAN-OCTAVIAN TRUICĂ received the B.Sc. degree in computer science and electrical engineering from the University Politehnica of Bucharest, Romania, in 2011, the B.Sc. degree in computer science and mathematics from the University of Bucharest, Romania, in 2013, and the M.Sc. degree in computer science engineering and information technology, and the Ph.D. degree in datamanagement and textmining from the University Politehnica of Bucharest, in 2013 and 2018, respectively. From 2019 to 2020, he was a Postdoctoral Researcher with the Data-Intensive Systems Group, Department of Computer Science, Aarhus University, Aarhus, Denmark, where he worked big data analytics for time series. During his Ph.D. studies, he was an Invited Researcher with the ERIC Laboratory, Université de Lyon, France, in 2015 and 2016, where he worked on data management, machine learning, and natural language processing. He is currently an Assistant Professor of computer science with the Department of Computer Science and Engineering, Faculty of Automatic Control and Computers, University Politehnica of Bucharest. His research interests include big data, data management, machine learning, text mining, natural language processing, and time series analysis.

ELENA-SIMONA APOSTOL received the Ph.D. degree from the University Politehnica of Bucharest, Romania, in 2014. She was a Postdoctoral Researcher with the Microsoft Research Center, Paris, in collaboration with the French Institute for Research in Computer Science and Automation (INRIA), where she worked on state-of-the-art big data analysis, multi-site cloud computing, and bioinformatics. She was an Invited Researcher during the Ph.D. studies at INRIA Rennes, France, working within the Joint Research Team between KerData at INRIA and UPB on big data management and analytics. During the bachelor's and master's studies, she was an Intern and a Junior Research Engineer with the Fraunhofer FOKUS Institute, Berlin, Germany, where she worked on computer networking and telecommunications with a focus on mobile and service orientated architectures. She is currently an Associate Professor of computer science with the Department of Computer Science and Engineering, Faculty of Automatic Control and Computers, University Politehnica of Bucharest (UPB). Her research interests include big data, data management, parallel and distributed algorithms, machine learning, and data science.