*Rev. Roum. Sci. Techn.– Électrotechn. et Énerg.*
Vol. **68**, *3*, pp. 295–300, Bucarest, 2023



# ROMANIAN TOPIC MODELING – AN EVALUATION OF PROBABILISTIC VERSUS TRANSFORMER-BASED TOPIC MODELING FOR DOMAIN CATEGORIZATION

MELANIA NIȚU[1], MIHAI DASCĂLU[1,2], MARIA-IULIANA DASCĂLU[3]

**Keywords: Topic modeling; Latent Dirichlet allocation (LDA); Bidirectional encoder representations from transformers topic (BERTopic); Clustering; Classification; Romanian documents.**

**When digitizing millions of volumes, a primary challenge for digital library systems is automatically analyzing and grouping the huge document collection by categories while identifying patterns and extracting the main themes. A common method to be leveraged on unlabeled texts is topic modeling. Given the wide range of datasets and evaluation criteria researchers use, comparing the performance and outputs of existing unsupervised algorithms is a complex task. This paper introduces a domain-based topic modeling evaluation applied to Romanian documents. Several variants of latent Dirichlet allocation (LDA) combined with dimensionality reduction techniques were compared to Transformer-based models for topic modeling. Experiments were conducted on two datasets of varying text lengths: abstracts of novels and full-text documents. Evaluations were performed against coherence and silhouette scores, while the validation considered classification and clustering tasks. Results highlighted meaningful topics extracted from both datasets.**

## 1. INTRODUCTION

A digital library's massive amount of text requires constant investment in automated processing tools to support fast access to relevant information. Topic modeling techniques frequently extract latent topics and concepts while grouping documents and words with similar meanings. Topic modeling considers unsupervised machine learning techniques with applications in Information Retrieval (IR) and Natural Language Processing (NLP) to support unlabeled textual data's comprehension, organization, and summarization.

In the digitization of millions of physical volumes, a priority is to review and organize the vast document collection by discovering hidden topical patterns across textual data, extracting key themes of the corpus, and using those insights to categorize documents and facilitate relevant document retrieval and transmission of meaning [1]. Topic modeling is one of the common approaches to be leveraged for this task and may represent an alternative to traditional recommender systems [2]. However, there are various topic modeling methods, whereas comparing algorithms' performance is a laborious process given the different evaluation criteria and datasets researchers employ. Thus, arguing for the accuracy and relevance of extracted topics is challenging.

This study performs a side-by-side comparison of state-of-the-art models, both statistical and deep learning-based, in combination with dimensionality reduction techniques. The experiments were conducted on two datasets of different text lengths: abstracts of novels and full-text documents. It is well known that short texts are generally more challenging to model topics as a lack of structure and noise often characterizes them. However, the corpus leveraged in this study needs to be revised. The short texts representing the abstracts of novels are manually written and curated by librarians. In contrast, the full texts were obtained via Optical Character Recognition (OCR), which outputs much noise, grammar errors, syntax issues, and lack of structure. This paper compares the results on both short and long Romanian text inputs against probabilistic and Transformer-based topic modeling.

### 1.1. TOPIC MODELING

This section examines state-of-the-art topic modeling algorithms. A common strategy for topic modeling is to consider a collection of term frequencies (TF), where the weight of each term additionally relies on the inverse document frequency (IDF) [3]. TF-IDF is most frequently computed as $f_{t,d} * \log\frac{N}{n_t}$, where $f_{t,d}$ is the frequency of term $t$ in document $d$, $N$ represents the number of documents, and $n_t$ is the number of documents where $t$ appears. The TF-IDF score increases proportionally with the importance of the word in the corpus.

Among statistical approaches, we mention latent Semantic analysis (LSA) [4], followed by the probabilistic latent semantic analysis (pLSA) [5], which was further developed into the latent Dirichlet allocation (LDA) [6]. LSA was proved more effective than pLSA and became one of the most popular methods.

LSA examines links between a group of documents and the terms included therein. It is mainly used for concept searching and automated document categorization. LSA applies singular value decomposition (SVD) to the term-document matrix to identify latent associations between concepts. Cells in the matrix contain the number of occurrences of that specific word in the document. LSA then applies a rank-lowering technique by measuring the fit between the data and the topic and then combines terms that have similar meanings.

The core idea behind pLSA is to determine the probability of certain terms being used with specific topics, analyze the co-occurrence matrix, and discover topics. It interprets topics as a probability distribution over words, with documents being a mixture of topics. The latent class model is decomposed through co-occurrences among words and documents.

[1] University Politehnica of Bucharest, Faculty of Automated Control and Computers, Splaiul Independentei 313, 060042, Bucharest, Romania
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044, Bucharest, Romania
[3] University Politehnica of Bucharest, Faculty of Engineering in Foreign Languages, Splaiul Independentei 313, 060042, Bucharest, Romania
E-mails: melania.nitu@yahoo.com, mihai.dascalu@upb.ro, maria.dascalu@upb.ro

DOI: 10.59277/RRST-EE.2023.68.3.8

LDA is a probabilistic approach based on the Bayesian approximation of posterior distributions; fundamentally, LDA can be perceived as the Bayesian version of pLSA. In LDA, each document is represented as a combination of latent topics, and each topic is modeled as a distribution across the words from the vocabulary. Additional extensions were developed to include continuous space word embeddings, applying multivariate Gaussian distributions on the embedding space instead of categorical distributions, thus resulting in its capability of handling out-of-vocabulary words. Gaussian LDA replaces the representation of discrete co-occurrence word counts with continuous embedded vectors.

From a statistical perspective, a few flavors of LDA were explored in this paper, such as Multicore and Mallet. Multicore is a parallelized streamed LDA that processes documents sequentially, speeding up the model training and making it ideal for large corpora. The sampling method represents the main difference between classic LDA and Mallet implementation. Standard LDA uses variational Bayes sampling, which is faster but less precise than Mallet's Gibbs sampling, relying on sampling from the conditional distributions of the target distribution.

Nonetheless, research shows that the probabilistic-based topic modeling algorithms require an investment in hyperparameter tuning to extract meaningful topics and an appropriate selection of evaluation metrics to assess extracted topics [7]. Additionally, topic models require an imposed number of topics, a custom stop words list, and pre-processing operations (i.e., stemming and lemmatization). Also, there are conditions when the models do not achieve good results – *e.g.*, when applied to a short text field [7].

Among the statistical approaches, Non-negative Matrix Factorization (NMF) is an unsupervised approach for factorizing matrices with non-negative values suitable for term-document matrices. It is a variant of SVD with additional restrictions for matrix decomposition imposed to resolve the issue of challenging interpretability, thus producing more interpretable and coherent topics [8].

Newer topic modeling techniques leverage NLP benchmark Transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers). The main difference between Transformer models and RNNs (recurrent neural networks) is their ability to be trained concurrently rather than sequentially. BERTopic [9] uses BERT to extract the document embedding, leveraging class-based TF-IDF and applying uniform manifold approximation (UMAP), described in the following section, to lower embedding dimensionality before clustering the documents using the density-based algorithm HDBSCAN [10]. HDBSCAN is a hierarchical method that handles irregular cluster shapes, identifies outliers, and outperforms former DBSCAN. Pre-trained Transformer-based models with BERT generate more accurate contextualized representations of words and sentences, thus supporting follow-up topic modeling.

Top2vec [11] is another Transformer-based topic modeling technique designed to address the inability to capture the semantics of words and documents. Top2vec leverages distributed representations of topics to identify topic vectors by using joint document and word semantic embedding. With this paradigm, the number of topics is automatically detected, and there is no need for stop words list, stemming, or lemmatization as opposed to probabilistic approaches. Also, this method can reduce the number of topics by merging the most similar topic vector; according to the authors, it produces more representative topics [11].

A distinct Transformer-based method is T-BERT [12], which enhances performance in sentiment classification by combining latent topics with contextual BERT embedding. T-BERT is a combination of LDA and BERT, which aims to obtain contextual topics on which the authors further apply BERT for sentiment analysis. The results are then clustered and visualized using the k-means clustering algorithm described in more detail in our Method section.

### 1.2. DIMENSIONALITY REDUCTION

Dimensionality reduction techniques are further applied to the output of topic modeling algorithms to reflect concepts instead of raw terms and provide a lower dimensional representation of documents, grouping words with similar semantics. In this study, we explore dimensionality reduction techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation (UMAP).

PCA [13] is a linear dimensionality reduction technique that provides an orthogonal projection of data in a lower dimensional space using the Gaussian distributions. In contrast, t-SNE [14] is a non-linear technique that measures the similarity between two data points by applying Student-t distributions and tries to group them while preserving the internal structure. UMAP [15] introduces optimizations for page management memory, making it suitable for data-intensive workloads. The model intends to keep as much of the global structure as possible and preserve the local structure. UMAP is based on three hypotheses that enable modeling as a fuzzy topological structure: the data is distributed uniformly on the Riemannian manifold, the Riemannian metric is constant, and the manifold is connected.

### 1.3. TOPIC MODELING FOR ROMANIAN

Studies on Romanian topic modeling are rare, but research shows promising results. For example, LDA and semantic recommendation techniques were combined to analyze Romanian literary life's chronology and capture the evolution of topics across historical periods [16].

Additionally, an analysis across seven languages, among which Romanian, was conducted in the Reminder European project [17] for comparative research across countries. The study aimed to understand how European migration topics are discussed similarly or differently in different languages and countries. LDA-based models were applied to a multilingual corpus, including 3,198 Romanian news articles published between 2014 and 2017. The findings show promising results with the polylingual topic model (PLTM) because the method enables the characterization of differences in topic prevalence at the document and language levels. However, to our knowledge, there are no baseline studies in the research field to evaluate topic modeling methods on Romanian corpora.

## 2. **METHOD**

### 2.1. CORPUS

Our core corpus consists of old Romanian documents, such as literary magazines and novels dated between the 19$^{th}$ century and the present, distributed in 17 domains, provided by

the Central University Library of Bucharest. Experiments were performed against two datasets with text of different lengths.

Corpus A is represented by 844 document descriptions manually written by librarians in the current form of the Romanian language, with a token distribution per document of around 0-500. Corpus B is represented by 635 full-text books, predominantly written in the 19th century using language terms specific to the respective era (*i.e.*, "aci" instead of "aici"), which were regulated in the text pre-processing stage and having a token distribution of around 0-500k words per document. Text pre-processing steps were performed on both datasets, such as text cleaning by using regular expressions to remove unwanted symbols and alphanumeric expressions mostly encountered after OCR (in the case of full-text documents), punctuation removal, stop words removal, tokenization, lemmatization, and noise reduction by eliminating words that are not in Romanian dictionary. Corpus dimension was considerably reduced after the processing steps: Corpus A lowered its dimension to around 0-200 tokens per abstract, while Corpus B was updated to 0-100 kwords per document.

We performed a corpus update for the validation phase. Corpus B did not have domain labeling, so we executed cross-correlation with Corpus A to replicate categories for common documents. This reduced corpus B from 635 to 247 documents only for validation purposes.

### 2.2. TOPIC MODELING METHOD

We conducted a set of experiments with different topic modeling techniques introduced in the first section, such as LDA in different configurations (default or standard, multicore, and Mallet) and BERT-based models (Sentence Transformer model and BERTopic), in combination with dimensionality reduction techniques, such as PCA, t-SNE, and UMAP.

Gensim implementation was leveraged for LDA standard, multicore, and the Mallet wrapper. The multicore configuration runs with constant memory. The maximum performance is reached when workers are set to the number of physical cores -1, keeping one core for the master process. The implementation of core estimation is based on Hoffman's research [18].

The number of optimal topics and parameter optimization was performed via grid search based on the coherence score (defined in the Evaluation metrics section). The highest values for coherence score (>70%) were achieved for $k = 10$ topics, *alpha* = asymmetric, and *beta* = 0.91.

BERT-based models leverage several options to extract document embeddings. First, we considered the sentence transformers (ST) [19] with an XML-R model that supports 50+ languages. For this study, the multilingual embedding model was leveraged in [19]. From documents, we create an array of sentences, eliminate stop words to reduce the noise, and return a vocabulary of sentences, which are encoded and served as input to the ST model. Second, the Flair framework was used to leverage Romanian BERT (*i.e.*, RoBERT [20]). We experimented with several pre-trained embedding models and compared results with the previously mentioned probabilistic techniques.

### 2.3. VISUALIZATION AND VALIDATION

Two different approaches were used to validate the results. Classification-based validation takes the extracted topics and tries to predict the documents' domains using a classification method. We considered several classification algorithms and evaluated how well the domains are predicted based on the extracted topics by reporting accuracy and F1 scores. Among the classification algorithms, we used a multi-layer perceptron (MLP) [21], extreme gradient boosting (XGBoost) (*i.e.*, an optimized distributed gradient-boosted decision trees) [22], Linear Regression (LR), and stochastic gradient descent (SGD). The Huber SGD was also considered, as this version smooths the loss and brings tolerance to outliers and probability estimates.

The second validation method was a cluster analysis using k-means [23], an iterative algorithm that partitions the dataset according to their features into k clusters. The extracted topics represent the dataset, and the optimal number of clusters is computed using the Elbow method [24]. This approach leverages WSS (Within the Sum of Squares) and the number of clusters to plot a curve; the inflection point gives the optimal number of clusters. A naïve observation is that the number of topics matches the number of optimal clusters. The resulting clusters are evaluated against purity, homogeneity, completeness, and v-measure. Purity measures the extent to which clusters contain a single class; each cluster is assigned a label based on the most frequent class, and purity is the number of correctly identified class and cluster labels divided by the number of total data points. Homogeneity evaluates the cluster labeling, given the ground truth. The completeness score measures if all data points of a given class are elements of the same cluster, while the v-measure represents the harmonic mean between homogeneity and completeness.

### 2.4. EVALUATION METRICS

Evaluating topic models is challenging due to their unsupervised nature and the absence of standardized measures. For this study, metrics such as coherence and Silhouette scores were leveraged.

Topic coherence is based on the hypothesis that words having comparable meanings are more likely to appear in a similar context and is computed using the Normalized Pointwise Mutual Information (NPMI) [25]. The NPMI score ranges between [-1; 1] and measures how closely the top ten words in a topic are linked to each other; a higher score means better coherence. Röder et al. [26] define coherence measure as a combination of segmentation of word subsets, probability estimation, confirmation measure, and aggregation. The $c_v$ coherence metric combines the indirect cosine measure with NPMI and the Boolean sliding window. Coherence score is computed by the formula described by

$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{\cos(\vec{w}_{n,k}, \vec{w}_k^*)}}{N \times K} \quad (1)$$

where c_v represents the average of all cosine similarities, $k$ is the topic index, $k \in \{1,2,\dots,K\}$, $n$ is the word index in a topic, $n \in \{1,2,\dots,N\}$, while $\vec{w}_{n,k}$ is the vector representing the topic word at index $n$ in topic $k$. For this study, Gensim's implementation of c_v was leveraged.

The Silhouette (S) coefficient also ranges from [-1;1] and measures the accuracy of the clustering technique. A higher score reflects distinguished clusters, while a score closer to the lowest limit means the clusters are wrongly assigned. A score close to 0 means the distance between clusters is insignificant. S is computed as $(b\text{-}a)/\max(a, b)$, where $a$ represents the average intra-cluster distance (*i.e.*, the average distance between each point within a cluster), and $b$ represents the average inter-cluster distance (*i.e.*, the average distance between all clusters).

## 3. RESULTS

Tables 1 and 2 introduce the coherence (c_v) and Silhouette (S) scores for corpus A and B, and the resulting S scores after applying different dimensionality reduction techniques such as PCA, t-SNE, and UMAP (i.e., marked as PCA_S, t-SNE_S, and UMAP_S). ST multilingual represents the ST model with multilingual embedding (distiluse-base-multilingual-cased). BERTopic (RoBERT) applies UMAP before clustering with HDBSCAN; therefore, no additional dimensionality reduction was applied.

*Table 1*

Evaluation of statistical versus Transformer-based topic models for Corpus A

| | c_v | S | PCA_S | t-SNE_S | UMAP_S |
|---|---|---|---|---|---|
| TF-IDF | 0.32 | 0.08 | 0.41 | 0.32 | 0.37 |
| LDA | | | | | |
| Default/Standard | 0.49 | 0.39 | 0.05 | 0.50 | 0.50 |
| Multicore | 0.54 | 0.53 | 0.44 | 0.23 | 0.20 |
| Mallet | 0.70 | 0.68 | 0.54 | 0.33 | 0.29 |
| BERT | | | | | |
| ST multilingual | 0.46 | 0.07 | 0.36 | 0.40 | 0.44 |
| BERTopic (Flair - RoBERT) | 0.69 | 0.71 | - | - | - |

*Table 2*

Evaluation of statistical versus Transformer-based topic models for Corpus B

| | c_v | S | PCA_S | t-SNE_S | UMAP_S |
|---|---|---|---|---|---|
| TF-IDF | 0.38 | 0.05 | 0.45 | 0.40 | 0.51 |
| LDA | | | | | |
| Default/Standard | 0.34 | 0.43 | 0.15 | 0.31 | 0.26 |
| Multicore | 0.52 | 0.51 | 0.24 | 0.21 | 0.17 |
| Mallet | 0.46 | 0.38 | 0.11 | 0.35 | 0.29 |
| BERT | | | | | |
| ST multilingual | 0.42 | 0.05 | 0.35 | 0.38 | 0.43 |
| BERTopic (Flair – RoBERT) | 0.82 | 0.69 | - | - | - |

Transformer models output better coherence than statistical methods. Previous studies demonstrated LDA does not work well on short texts; however, in our case, coherence was better for Corpus A, while Transformer-based models (BERTopic) outperformed Corpus B. This behavior can be argued because Corpus A contains cleaned text (*i.e.*, manually written and corrected by librarians). In contrast, corpus B contains text automatically extracted via OCR from scanned old documents, which introduces many noise and spelling errors. Even though the text was preprocessed to reduce the syntax issues, Corpus A has a superior quality to Corpus B.

Nonetheless, the study by Chang et al. [27] shows that statistical methods can only partially reflect the human perception of topic coherence; additional human judgment is often required to assess the extracted topics. This represents a qualitative assessment based on human interpretability and understanding. Figure 1 illustrates the top 5 words within extracted topics in corpus B using BERTopic, which achieved very good results among the Transformer experiments (*i.e.*, 82.09% c_v on corpus B). Each topic presents meaningful associations between its terms. As exposed by Topic 3, the term "nicolae_iorga" represents a Romanian author having the most important contribution to our dataset, authoring around 10% of the texts in corpus B. His name is also present in other writings as an influential figure of the 19th century.

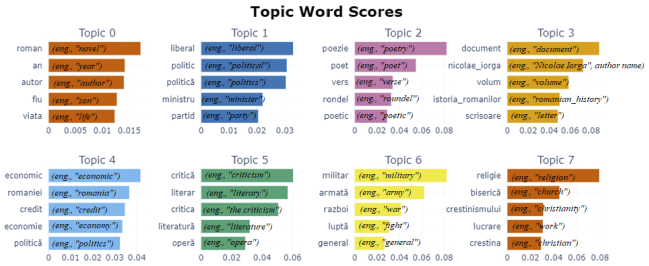



Figure 1. Overview of BERTopic extracted *topics*.

Validations were performed from two perspectives: cluster quality and domain-based classification, which essentially checks if categories are accurately identified based on extracted topics.

*Table 3*

Validation using classification technique for LDA Mallet and BERTopic (BT) on corpus A (abstract) and corpus B (all text)

| | | MLP | | XGBoost | | LR | | SGD | | SGD Huber | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LDA | BT | LDA | BT | LDA | BT | LDA | BT | LDA | BT |
| Accuracy | A (abstract) | .87 | .44 | .82 | .53 | .68 | .43 | .52 | .39 | .55 | .43 |
| | B (all text) | .55 | .60 | .78 | .61 | .56 | .40 | .85 | .34 | .87 | .32 |
| F1-macro avg | A (abstract) | .86 | .09 | .82 | .21 | .69 | .06 | .52 | .05 | .54 | .06 |
| | B (all text) | .55 | .16 | .77 | .18 | .55 | .06 | .81 | .05 | .86 | .04 |
| F1-weighted avg | A (abstract) | .87 | .33 | .81 | .47 | .67 | .31 | .52 | .26 | .54 | .31 |
| | B (all text) | .55 | .37 | .78 | .30 | .56 | .27 | .85 | .20 | .87 | .18 |

Accuracy and F1 scores for classification-based validations are presented in Table 3. The dataset was split into 60 % train and 40 % test samples for validation. Mallet achieved an accuracy score of 87 % on Corpus A with the MLP classifier, and the same score was reached with SGD on Corpus B, while BERTopic scores best with XGBoost (53 % on Corpus A and 61 % on Corpus B). As an observation, in this experiment, the coherence score is not proportional to the accuracy or the F1 score. It is well known that BERT models perform better when a higher quantity of text is fed as input to the neural network. Therefore, the expectation is that corpus B will produce better results with BERT than with LDA. The hypothesis is partially verified with MLP validation, while the results are counterintuitive for the others. This can be explained by the poor quality of corpus B (noisy, OCRed text) compared to corpus A, which contains well-written text.

*Table 4*

Validation using clustering technique for LDA and BERTopic(BT) on corpus A(abstract) and corpus B(all text)

| Metric | Corpus | LDA | BT |
|---|---|---|---|
| Cluster Purity | A (abstract) | .40 | .48 |
| | B (all text) | .51 | .56 |
| Homogeneity | A (abstract) | .11 | .21 |
| | B (all text) | .39 | .25 |
| Completeness | A (abstract) | .09 | .21 |
| | B (all text) | .36 | .28 |
| V-measure | A (abstract) | .10 | .21 |
| | B (all text) | .37 | .25 |

The validation via the clustering technique is presented in Table 4, where we assessed the quality of the clusters. According to the validation results on the current corpus, we

can better predict document domains with LDA rather than with BERT-extracted topics. This can be justified by the low number of documents and the fact that Transformer models exhibit better results when working with large collections.

An additional validation leverages the Elbow method to identify the optimal number of clusters. A naïve observation is that the number of topics matches the number of optimal clusters. The clustering results are presented in Fig. 2.

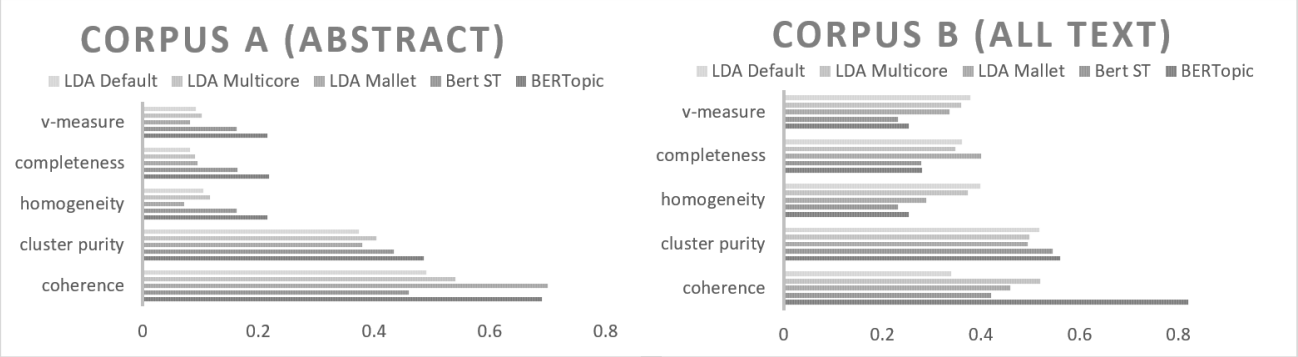
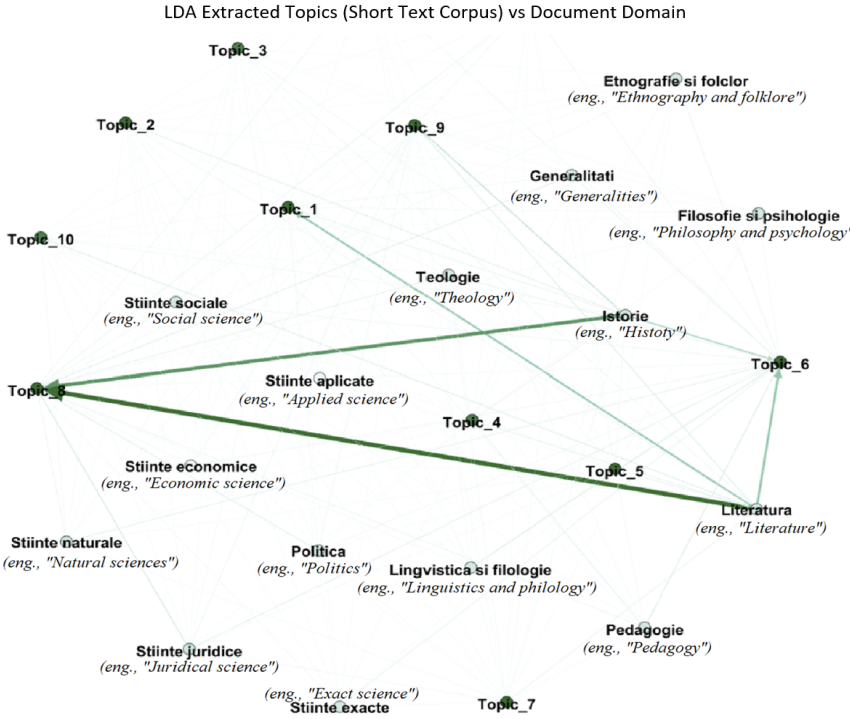



Fig. 2 – Cluster evaluation against extracted topics.

The coherence score was listed to check whether there is a correlation between coherence and cluster metrics. The results illustrate that coherence is not proportional to cluster purity, like the first validation method. Better performance is achieved for Corpus B; as such, the amount of text (*i.e.*, dataset size) directly influences the quality of the clusters for extracted topics.

The human assessment was facilitated by exposing the extracted topics in an interactive graphic using the PyLDAvis visualization tool. Topics are plotted in a two-dimensional space by computing the distance between concepts and using multidimensional scaling to project inter-topic distances onto two dimensions. The top 10 topics were extracted via LDA with relevance metric $\lambda = 0.59$, determining the term-topic specificity. The relevance metric was tuned to ensure correct specificity and term coherence.




| Topic_1 | Topic_2 | Topic_3 | Topic_4 | Topic_5 | Topic_6 | Topic_7 | Topic_8 | Topic_9 | Topic_10 |
|---|---|---|---|---|---|---|---|---|---|
| opera (eng.,"opera") | militar (eng.,"military") | negustor (eng.,"merchant") | traduce (eng.,"translate") | genealogic (eng.,"genealogic") | lucrare (eng.,"work") | drept (eng.,"law") | document (eng.,"document") | mare (eng.,"great") | piesa (eng.,"piece") |
| literar (eng.,"literary") | război (eng.,"war") | medieval (eng.,"medieval") | cercetător (eng.,"researcher") | suflet (eng.,"soul") | autor (eng.,"author") | corespondenta (eng.,"mail") | volum (eng.,"volume") | politica (eng.,"politics") | dramatic (eng.,"dramatic") |
| scrisoare (eng.,"letter") | religie (eng.,"religion") | seminar (eng.,"seminar") | pedagogie (eng.,"pedagogy") | munci (eng.,"work") | roman (eng.,"novel") | noțiuni (eng.,"notions") | an (eng.,"year") | lucrare (eng.,"work") | dramatic (eng.,"dramatic") |
| critica (eng.,"criticism") | armata (eng.,"army") | veniamin (eng.,"veniamin") | copil (eng.,"kid") | animal (eng.,"animal") | nicolae_iorga (eng.,"nicolae_iorga") | univers (eng.,"univers") | istoric (eng.,"historian") | stat (eng.,"state") | tragedie (eng.,"tragedy") |
| idee (eng.,"idea") | cronica (eng.,"chronicle") | orașelor (eng.,"cities") | rol (eng.,"role") | corp (eng.,"body") | studio (eng.,"studio") | serviciu (eng.,"opera") | roman (eng.,"novel") | basarabia (eng.,"bessarabia") | teatru (eng.,"theater") |
| metoda (eng.,"method") | lupta (eng.,"fight") | mediu (eng.,"medium") | educației (eng.,"education") | practica (eng.,"practical") | parte (eng.,"part") | orașe (eng.,"cities") | act (eng.,"document") | parte (eng.,"part") | morții (eng.,"deaths") |
| literatura (eng.,"literature") | războiului (eng.,"of the war") | inscripții (eng.,"inscriptions") | învățământului (eng.,"education") | rural (eng.,"rural") | volum (eng.,"volume") | didactic (eng.,"didactic") | mare (eng.,"great") | problema (eng.,"problem") | scena (eng.,"scene") |
| viață (eng.,"life") | sanitar (eng.,"sanitary") | poesia (eng.,"poem") | manuscris (eng.,"manuscript") | patriotic (eng.,"patriotic") | limba (eng.,"language") | matematica (eng.,"mathematics") | autor (eng.,"author") | tare (eng.,"loud") | greva (eng.,"strike") |
| spirit (eng.,"spirit") | programa (eng.,"program") | carte (eng.,"book") | învățământului (eng.,"education") | intelectual (eng.,"intellectual") | istorie (eng.,"history") | institut (eng.,"institute") | istorie (eng.,"history") | autor (eng.,"author") | acțiunea (eng.,"action") |
| lume (eng.,"world") | folos (eng.,"use") | lirica (eng.,"lyrical") | elev (eng.,"student") | două_jumătate (eng.,"two_half") | literatura (eng.,"literature") | anterior (eng.,"previous") | lucrare (eng.,"work") | boier (eng.,"lord") | episcop (eng.,"bishop") |

Fig. 3 – Correlation domain – topics

In addition, the Gephi tool [28] was leveraged to plot the domain-topics matrix and check if the extracted topics are

correctly associated with document domains (see Fig. 3). It displays a clear correlation between the extracted topics and the domain of documents. The resulting graph exposes a strong correspondence that matches human judgment, and the extracted terms belong to the same semantic field. Such examples are detailed below.

- "literatură" (*eng., "literature"*) – topic 1, topic 6
- "istorie" (*eng.,* "*history"*) – topic 2, topic 9
- "științe juridice" (*eng.,* "*juridical science"*)– topic 7
- "pedagogie" (*eng., "pedagogy"*) – topic 4

## 4. **CONCLUSIONS**

This paper introduced an evaluation of topic modeling algorithms, both probabilistic and Transformer-based, against Romanian corpora of different text lengths (Corpus A versus Corpus B). The Transformer-based technique scored top coherence for topics extracted via BERTopic leveraging the RoBERT model (69 % coherence on corpus A and 82 % on corpus B). In comparison, LDA scored a maximum of 70% on corpus A for the Mallet implementation and 62% on corpus B for the multicore implementation.

Dimensionality reduction techniques did not improve performance; we may observe meaningful results in some cases by using human judgment to analyze the resulting topics. Since we got smaller coherence for short text, we further investigated the output using Gephi and observed a high correlation for domain-topics mappings.

Additional validations were performed using classification-based and clustering techniques. We successfully predicted the documents' domains starting from the extracted topics and an MLP with an accuracy of 87 % for the LDA-based approach on corpus A. The same score was reached on corpus B using SGD validation. For the BERTopic approach, we scored an accuracy of 53 % on corpus A and 61 % on corpus B via the XGBoost classification technique. This argued that topics were meaningful for the respective categories. Moreover, the smaller accuracy obtained using BERTopic is an expected behavior because of the poor quality of corpus B and the small quantity of text for validation corpus. In contrast, the clustering validation shows better results for LDA on corpus B, whereas BERTopic scored better on corpus A.

While analyzing the overall results, we are aligned with previous research [27] that argued that the coherence score could be better for human perception. Validation shows that a high coherence does not imply the extracted topics are always relevant, according to the accuracy results.

## ACKNOWLEDGMENT

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS –UEFISCDI, project number TE 151 from 14/06/2022, within PNCDI III: "Smart Career Profiler based on a Semantic Data Fusion Framework".

*Received on 15 February 2023*

## REFERENCES

1. R. Dobrescu, D. Merezeanu, *From information to knowledge transmission of meaning*, Rev. Roum. Sci. Techn., **62**, *1*, pp. 115–118 (2017).
2. R.-I. Mogoș, C.-N. Bodea, *Recommender systems for engineering education*, Rev. Roumn. Sci. Tech., **64**, *4*, pp. 435–442 (2019).
3. A. Rajaraman, J.D. Ullman, *Data Mining: Mining of Massive Datasets*, pp. 1–17 (2011).
4. S.T. Dumais, *Latent semantic analysis*, Annual Review of Information Science and Technology, **38**, *1*, pp. 188–230 (2004).
5. T. Hoffman, *Unsupervised learning by probabilistic latent semantic analysis*, Machine Learning, **42**, *1*, pp. 177–196 (2001).
6. D.M. Blei, A.Y. Ng, M.I. Jordan, *Latent Dirichlet allocation*, Journal of Machine Learning Research, **3**, *4-5*, pp. 993–1022 (2003).
7. L. Hong, B.D. Davison, *Empirical study of topic modeling in Twitter*, ACM, pp. 80–88 (2010).
8. N. Gillis, *The why and how of nonnegative matrix factorization*, arXiv:1401.5226 (2014).
9. M. Grootendorst, *BERTopic: neural topic modeling with a class-based TF-IDF procedure*, arXiv:2203.05794 (2022).
10. L. McInnes, J. Healy, *Accelerated hierarchical density based clustering*, ICDMW, pp. 33–42 (2017).
11. D. Angelov, *Top2Vec: distributed representations of topics*, arXiv:2008.09470 (2020).
12. S. Palani, P. Rajagopal, S. Pancholi, *T-BERT - model for sentiment analysis of micro-blogs integrating topic model and BERT*, arXiv:2106.01097 (2021).
13. C.M. Bishop, *Pattern recognition and machine learning*, New York, Springer (2006).
14. L. van der Maaten, G. Hinton, *Visualizing data using t-SNE*, Journal of Machine Learning Research, **9**, pp. 2579–2605 (2008).
15. L. McInnes, J. Healy, *UMAP: uniform manifold approximation and projection for dimension reduction*, arXiv:1802.03426 (2018).
16. L.-M. Neagu, T.-M. Cotet, M. Dascalu, S. Trausan-Matu, E. Chisu, E. Simion, *Semantic recommendations and topic modeling based on the chronology of Romanian literary life*, SETE, pp. 164–174 (2019).
17. F. Lind, J.-M. Eberl, S. Galyga, T. Heidenreich, H.G. Boomgaarden, B.H. Jimenez, R., Berganza, *A bridge over the language gap: topic modelling for text analyses across languages for country comparative research*, REMINDER project (2019).
18. M.D. Hoffman, D.M. Blei, F. Bach, *Online learning for latent Dirichlet allocation,* NIPS, pp. 856–864 (2010).
19. N. Reimers, I. Gurevych, *Sentence-BERT: sentence embeddings using siamese BERT-networks*, EMNLP-IJCNLP, pp. 3982–3992 (2019).
20. M. Masala, S. Ruseti, M. Dascalu, *RoBERT-a Romanian BERT model*, COLING, pp. 6626–6637 (2020).
21. F. Murtagh, *Multilayer perceptrons for classification and regression*, Neurocomputing, **2**, pp. 183-197 (1991).
22. T. Chen, C. Guestrin, *XGBoost: a scalable tree boosting system*, KDD, pp. 785–794 (2016).
23. J.A. Hartigan, M.A. Wong, *A k-means clustering algorithm*, RSS, **28**, *1*, pp. 100-108 (1979).
24. R.L. Thorndike, *Who belongs in the family?*, Psychometrika, **18**, pp. 267–276 (1953).
25. G. Bouma, *Normalized (pointwise) mutual information in collocation extraction* (2009).
26. M. Röder, A. Both, A. Hinneburg, *Exploring the space of topic coherence measures*, WSDM, pp. 399–408 (2015).
2.7 J. Chang, S. Gerrish, W. Chong, J.L. Boyd-Graber, D.M., Blei, *How humans interpret topic models*, NeurIPS, pp. 228-296 (2009).
28. M. Bastian, S. Heymann, M. Jacomy, *Gephi: An open source software for exploring and manipulating networks*, AAAI ICWSM, pp. 361–362 (2009).