

Uncovering Latent Scientific Structures in Albanian Texts Using Transformer-Based Clustering

LEONA REXHAJ, LUELA PRIFTI

Department of Mathematical Engineering,

Polytechnic University of Tirana, Faculty of Mathematical Engineering and Physics Engineering
ALBANIA

Abstract: - This study analyzes the semantic structure of an Albanian corpus comprising scientific texts with 11.5 million words and approximately 74,500 paragraphs. It aims to cluster the documents according to semantic similarity and uncover thematic structures. The proposed methodology integrates Sentence-BERT to construct semantic representations, UMAP for dimensionality reduction, and HDBSCAN for clustering. Large-scale analysis was conducted without predefined labels, allowing the clusters to emerge directly from the textual content. The paraphrase-multilingual-MiniLM-L12-v2 model achieved the best clustering performance and produced a stable seven-cluster structure. Two representation strategies were evaluated: full-text representation and section-based representation using abstracts, introductions, and conclusions. External validation metrics for both representations indicated moderate alignment with disciplinary classifications, reflecting the interdisciplinary nature of scientific texts. For the section-based representation, the Silhouette score was 0.7154, the Davies-Bouldin index was 0.5025, and the Calinski-Harabasz index was 651.99. The relationship between the resulting clusters and academic fields was statistically significant, with $\chi^2 = 748.443$ and $p = 4.90 \times 10^{-134}$, while Cramer's $V = 0.746$ indicated a strong association. These results suggest that the section-based representation produces clearer and more semantically coherent clusters by focusing on the main thematic information and reducing non-discriminative content. The proposed framework offers a practical and scalable approach for analyzing large scientific corpora and may be adapted to other languages and low-resource research domains.

Key-Words: - Semantic modeling; Transformer-based embeddings; Sentence-BERT; Unsupervised clustering; UMAP; HDBSCAN; Albanian corpus; Text mining; Low-resource NLP.

Received: March 25, 2026. Revised: May 20, 2026. Accepted: June 23, 2026. Published: September 18, 2026.

1 Introduction

Nowadays, the development of Natural Language Processing (NLP) has accelerated significantly with the use of transformer models. By considering the context in which words are used, these models make it possible to represent the meaning of a text more effectively. They have been applied to various tasks, including document clustering, topic modeling, and content analysis, although most studies continue to focus on high-resource languages, particularly English [1], [2], [3]. Albanian remains relatively underexplored in large-scale semantic analysis [4]. The Albanian research in the field of Natural Language Processing is focused mainly on developing fundamental resources and tools, including corpora for named entity recognition [5], morphological tagging and lemmatization [6], datasets for topic modeling [7], and supervised classification tasks [8]. However, comprehensive approaches to unsupervised semantic analysis remain limited.

The analysis of scientific texts is even more challenging because they use specialized terminology, address a wide range of topics, and often combine concepts from several disciplines. Traditional methods such as bag-of-words and TF-IDF are limited in capturing deep semantic relationships, whereas transformer-based embedding models such as Sentence-BERT (SBERT) [1] represent documents in a continuous semantic space [9]. Embedding-based approaches combined with dimensionality reduction and clustering algorithms such as UMAP and HDBSCAN have been shown to uncover latent structures in large text corpora [2], [3], [10], [11], [12]. Recent studies on transformer-based classification and similarity-distance modeling [13], [14] further support the relevance of representation-based methods for structured text analysis. However, such methodologies have rarely been evaluated in low-resource languages [4] and have not yet been applied to large-scale Albanian scientific corpora.

This study investigates a large corpus of Albanian scientific texts using an unsupervised semantic analysis framework to model the underlying semantic space and uncover latent thematic structures without predefined labels [10], [15], [16]. The main contribution of this work is the systematic comparison of full-text and section-based representation strategies, where abstracts, introductions, and conclusions are considered semantically dense components of scientific documents. The proposed framework is evaluated using clustering performance metrics, stability analysis, and statistical tests of cluster-discipline associations. A key methodological contribution is the development of a reproducible framework that integrates semantic representation, dimensionality reduction, clustering, and statistical validation. Rather than performing an exhaustive comparison of alternative methods, such as BERTopic, spectral clustering, graph-based approaches, and contrastive sentence embeddings, this work focuses on the influence of document representation design. Within the same SBERT-UMAP-HDBSCAN framework, full-text and section-based representations are compared, allowing the impact of representation strategy on clustering performance, semantic coherence, and interpretability to be systematically analyzed.

2 Methodology

The proposed methodology provides a scalable, unsupervised framework for semantic analysis of scientific texts in a low-resource language setting [4]. It combines multilingual transformer-based embeddings with nonlinear dimensionality reduction and density-based clustering to uncover latent thematic structures without relying on annotated data. The proposed pipeline consists of three main stages: semantic representation using SentenceBERT [1], dimensionality reduction via UMAP [15], and clustering using HDBSCAN [10], [16].

2.1 Corpus Description and Preprocessing

The dataset consists of Albanian-language scientific texts and contains a total of 11,504,932 words organized into 74,527 paragraphs. This large amount of material provides a suitable basis for conducting embedding-based semantic analysis. The documents vary considerably in length: the shortest contains 9,638 words, whereas the longest reaches 187,040 words. The mean document length is 47,937 words, the median is 44,376, and the standard deviation is 23,187 words. Such variability

may bias representation learning and affect the stability of density-based clustering. Preprocessing includes normalization steps such as removal of redundant whitespace and reference sections. The documents are divided into 200-word segments to maintain semantic consistency and minimize the influence of differences in document length on the analysis.

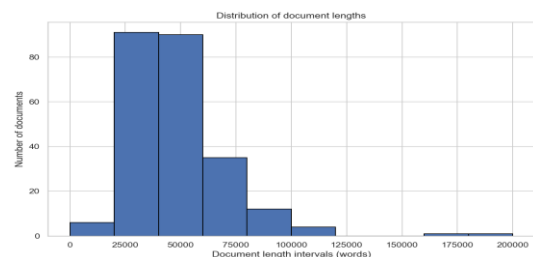


Fig. 1: Document length distribution

Source: created by the authors

Figure 1 illustrates how the documents are distributed by length. Most documents contain between 20,000 and 60,000 words, while their frequency decreases considerably beyond 70,000 words. Only a small number exceed 100,000 words, and the longest documents approach 200,000 words. This indicates that the corpus is mainly composed of medium-length documents, whereas a limited number of very long documents account for a considerable share of its total volume.

The number of paragraphs also varies substantially from one document to another. On average, each document contains approximately 310 paragraphs, while the median is 275 paragraphs and the standard deviation is 165 paragraphs. These values indicate that the documents do not follow a uniform structure.

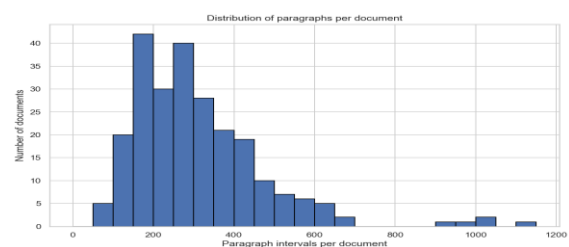


Fig. 2: Number of paragraphs per document

Source: created by the authors

As shown in Figure 2, the distribution of paragraphs per document is right-skewed, with most documents concentrated between 200 and 400 paragraphs. However, the number of paragraphs varies considerably from one document to another. In some cases, documents contain more than 800 paragraphs, while the maximum reaches approximately 1,100, indicating that certain texts are

highly segmented. Regarding paragraph length, each paragraph contains an average of 155 words. The median is 138 words, while the standard deviation is 72 words.

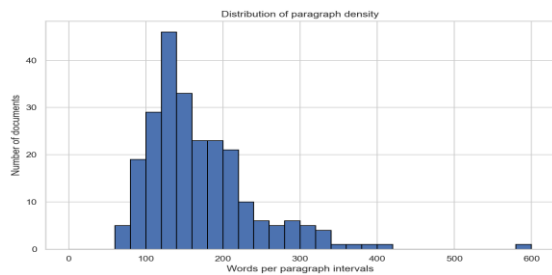


Fig. 3: Paragraph lengths
Source: created by the authors

Figure 3 presents the distribution of paragraph lengths, which is more concentrated than the previous distributions but remains slightly right-skewed. Most paragraphs contain between 100 and 180 words, while only a few exceed 300 words and one approaches 600 words. The interquartile range (IQR) method applied to document lengths identifies eight unusually long documents containing 96,336; 99,028; 108,186; 114,241; 117,540; 119,121; 176,771; and 187,040 words. Although limited in number, these unusually long documents contribute to the heavy-tailed distribution and may influence semantic modeling by giving longer, information-dense texts greater weight in the semantic space. To reduce this effect, segment aggregation is used, allowing documents to contribute more evenly and producing more stable and coherent embeddings.

2.2 Semantic Representation

Let $D = \{d_1, d_2, \dots, d_n\}$ denote the document collection. Each document d_i is represented in a continuous vector space, where semantic similarity is reflected through geometric proximity. Since the documents are long and structurally heterogeneous, each document d_i is partitioned into k_i segments. Segment-level embeddings are then computed using Sentence-BERT [1]. The document-level embedding is obtained according to Equation (1):

$$\mathbf{v}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbf{e}_{ij} \quad (1)$$

where $\mathbf{v}_i$ denotes the embedding of document d_i , $\mathbf{e}_{ij}$ denotes the embedding of its j -th segment, and k_i is the number of segments in the document d_i . This averaging-based segment aggregation represents each document as the centroid of its segment-level semantic vectors. Segment aggregation is an established unsupervised strategy for semantic representation, particularly when local variations

and dominant common components are controlled [17]. In the present framework, segment aggregation is used to reduce local semantic variations introduced by examples, formulas, technical details, or supplementary explanations while preserving the dominant thematic direction of each document. It also reduces the effect of document length, since each document is represented by a single vector. The document-level embeddings are L2-normalized before dimensionality reduction, and clustering is applied.

A comparative evaluation is conducted using three pretrained embedding models: paraphrase-multilingual-MiniLM-L12-v2, multilingual-e5-large, and all-MiniLM-L6-v2. These models are selected because of their established use in multilingual semantic representation tasks [1], [9], [11], [12]. Truncated Singular Value Decomposition (TruncatedSVD) [18] is applied to reduce dimensionality and improve computational efficiency, after which UMAP is used to construct the reduced UMAP space for clustering.

2.3 Dimensionality Reduction

Dimensionality reduction is performed using UMAP (Uniform Manifold Approximation and Projection) [15], a nonlinear technique designed to preserve local neighborhood relationships in high-dimensional spaces. Alternative dimensionality reduction approaches, such as t-SNE [19] and multidimensional scaling [20], were not adopted because they are less suitable for large-scale density-based clustering. UMAP is configured with $n_neighbors = 9$, $min_dist = 0.01$, $n_components = 12$, and cosine distance as the metric. Although exhaustive hyperparameter optimization is beyond the scope of this study, multiple UMAP initializations show that the dominant seven-cluster structure remains stable.

2.4 Clustering

Clustering is performed using HDBSCAN [10], [16]. This algorithm can identify clusters with different densities while also handling data points that do not clearly belong to any cluster in complex, nonlinear spaces. For grouping the documents under study, HDBSCAN is configured with $min_cluster_size = 12$ and $min_samples = 2$, while Euclidean distance is selected as the distance metric. These parameters are selected to identify compact and interpretable clusters while allowing semantically ambiguous or boundary documents to be assigned the noise label. UMAP measures similarity between embeddings using cosine distance in the high-dimensional space, while

HDBSCAN applies Euclidean distance in the reduced UMAP space. Since HDBSCAN determines the number of clusters automatically, the combined UMAP-HDBSCAN framework is suitable for semantic data with overlapping cluster boundaries.

2.5 Representation Strategies

Two representation strategies are evaluated: full-text representation and section-based representation. The full-text representation uses the complete document content, preserving the overall informational context but potentially introducing semantic noise from formulas, technical descriptions, examples, and peripheral explanations. Such noise may increase intra-document variability and reduce cluster separability. In contrast, the section-based representation focuses on abstracts, introductions, and conclusions, which typically contain higher-level research information and greater thematic concentration. This representation is expected to reduce irrelevant semantic variation and provide a more discriminative basis for improving cluster compactness, separability, and interpretability.

2.6 Evaluation Metrics

Clustering performance is assessed using internal and external metrics, together with statistical validation measures. Three internal metrics are used to evaluate the results: the Silhouette Score [21], the Davies-Bouldin Index [22], and the Calinski-Harabasz Index [23]. Higher Silhouette and Calinski-Harabasz values generally indicate more compact clusters with clearer separation. In the case of the Davies-Bouldin Index, a lower value indicates better separation between clusters. External validation measures include Purity and Normalized Mutual Information (NMI), which quantify correspondence with predefined disciplinary categories, and the Adjusted Rand Index (ARI) [24], which measures agreement between partitions while correcting for chance. The chi-square test and Cramer's V are used to assess the statistical association between clusters and academic fields. Clustering uncertainty is assessed through noise assignments and cluster stability under different UMAP initializations. HDBSCAN identifies observations outside sufficiently dense regions as noise, indicating boundary or low-confidence cases. The stability of this uncertainty indicator is evaluated by comparing noise assignments across the tested initializations. Accordingly, the resulting clusters are interpreted as approximations of latent thematic categories rather than predefined disciplinary labels.

2.7 Theoretical and Computational Aspects

The proposed framework can be interpreted as a staged approximation of the latent semantic structure of the corpus. At the representation level, each segment embedding is given by Equation (2):

$$\mathbf{e}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij} \quad (2)$$

where $\boldsymbol{\mu}_i$ denotes the dominant thematic component of document d_i , and $\boldsymbol{\epsilon}_{ij}$ represents local semantic variation introduced by examples, formulas, technical details, or peripheral explanations. This approximation assumes that local semantic perturbations are weakly dependent, approximately zero-mean, and have finite variance. Under the ideal assumption of independence, the variance of the sample mean decreases proportionally to $1/k_i$. Therefore, segment aggregation provides a more stable document-level embedding than direct full-text embedding. This effect stabilizes neighborhood relations and reduces groupings driven by local textual noise. The section-based representation further supports this objective by selecting text with higher thematic density, such as abstracts, introductions, and conclusions. These sections contain more discriminative semantic information and less auxiliary content, increasing the semantic signal-to-noise ratio before dimensionality reduction and clustering. From a computational perspective, preprocessing and segmentation scale linearly with the number of textual units. Embedding generation depends on the number of segments and the transformer encoding cost. UMAP first organizes the data by considering the nearest neighbors and then represents them in a lower-dimensional space. Within this reduced space, HDBSCAN identifies regions where the data form stable density-based clusters. Using HDBSCAN after UMAP improves the efficiency of the clustering process, as the algorithm operates on a lower-dimensional representation of the data, which is particularly beneficial for large datasets.

Table 1. Computational complexity of the proposed semantic clustering framework

Pipeline stage	Approximate complexity
Preprocessing and segmentation	$O(N)$
Segment-level embedding generation	$O(S \cdot C_{enc})$
Segment aggregation	$O(Sd)$
UMAP dimensionality reduction	$O(n \log n)$
HDBSCAN clustering	$O(n \log n)$ to $O(n^2)$

Source: created by the authors based on [1], [10], [15], and [16]

In Table 1, N denotes the total number of words processed during preprocessing and segmentation, S the number of generated segments, d the embedding dimension, n the number of documents, and C_{enc} the computational cost of transformer-based encoding. The stated complexities are approximate and may vary depending on implementation details and data structure.

3 Results and Discussion

This section evaluates the ability of the proposed unsupervised framework to uncover latent thematic structures in Albanian scientific texts. The analysis considers embedding model performance, clustering stability, cluster structure, and the impact of representation strategies on clustering performance and interpretability. Results are compared within a fixed analytical pipeline, while robustness is assessed across multiple UMAP initializations.

3.1 Model Comparison and Selection

To identify the most suitable model for semantic representation of Albanian scientific texts, a comparative analysis is conducted across three transformer-based models: *paraphrase-multilingual-MiniLM-L12-v2* (M1), *multilingual-e5-large* (M2), and *all-MiniLM-L6-v2* (M3). Each of the three models is tested using the same analytical procedure and the segment aggregation method presented in Equation (1). This allows for the comparison of their results under the same conditions. Previous studies have shown that transformer-based embeddings can provide stable semantic representations across different languages [1]. Table 2 presents the clustering performance metrics for each model.

Table 2. Comparison of clustering performance across models

Metric / Model	M1	M2	M3
Dimension	384	1024	384
Silhouette (Silh.)	0.5349	0.5012	0.4726
Davies-Bouldin (DB)	0.7608	0.8214	0.9037
Calinski-Harabasz (CH)	219.37	201.58	176.42

Source: created by the authors

Among the three models tested, paraphrase-multilingual-MiniLM-L12-v2 produces the best results across all internal metrics, achieving the highest Silhouette value and the lowest Davies-Bouldin value. In contrast, multilingual-e5-large does not improve clustering despite using higher-dimensional embeddings, showing that greater

dimensionality does not necessarily lead to clearer cluster separation. The weakest results are obtained with all-MiniLM-L6-v2, which produces less compact and less clearly separated clusters. These findings suggest that embedding regularity, neighborhood consistency, and semantic compactness are more important than dimensionality alone in unsupervised clustering. The stronger performance of paraphrase-multilingual-MiniLM-L12-v2 may be attributed to a more stable and discriminative embedding space, whereas higher-dimensional embeddings may introduce additional variability that does not necessarily improve cluster separation. Based on this analysis, paraphrase-multilingual-MiniLM-L12-v2 is selected for the subsequent experiments.

3.2 Clustering Stability Analysis

The robustness of the clustering structure is assessed by examining the impact of different UMAP initializations [15]. Table 3 reports clustering outcomes across multiple UMAP initializations.

Table 3. Clustering stability across UMAP initializations

Random state	Clusters	Noise	Silh.	DB	CH
0	7	10	0.532	0.768	217.9
1	7	11	0.529	0.776	214.6
7	7	9	0.537	0.758	221.4
21	6	12	0.498	0.846	191.7
42	7	10	0.535	0.761	219.37
100	8	13	0.507	0.821	203.5
123	7	9	0.534	0.761	218.7
202	7	10	0.531	0.770	216.2
999	7	11	0.528	0.778	213.9

Source: created by the authors

A seven-cluster structure is obtained in seven of the nine runs, with consistently favorable internal validation metrics. The Silhouette and Calinski-Harabasz values are lower, while the Davies-Bouldin value is higher for the runs producing six or eight clusters. This indicates that the resulting clusters are less compact and less clearly separated from one another. Small variations in the number of documents classified as noise do not significantly affect the overall clustering structure. These noise-assigned documents may correspond to semantically ambiguous or boundary cases and provide a practical indication of clustering uncertainty. The results indicate that the identified thematic structure is robust across the tested UMAP initializations.

3.3 Clustering Performance Evaluation

Clustering performance is evaluated using the internal and external metrics described in Section 2.6.

Table 4. Clustering performance metrics

Metric	Value
Purity	0.63
NMI	0.3968
ARI	0.35
Silhouette	0.5349
Davies-Bouldin	0.7608
Calinski-Harabasz	219.37

Source: created by the authors

The results in Table 4 indicate a coherent thematic structure. This interpretation is also supported by internal evaluation metrics. The Silhouette score of 0.5349 points to a reasonably cohesive cluster structure, while the Davies-Bouldin value of 0.7608 shows that the clusters are separated to a satisfactory degree. The Calinski-Harabasz value of 219.37 also suggests that the clustering structure is compact [21], [23]. The external metrics show moderate alignment with disciplinary classifications. The values of 0.63 and 0.3968, respectively, for Purity and NMI indicate meaningful thematic grouping, while the ARI value of 0.35 reflects partial agreement with predefined categories [24]. We expected this moderate external agreement because the predefined academic-field labels are broad, whereas clustering captures latent semantic relationships across disciplines. The combination of internal cohesion, external agreement, and interpretability supports the presence of meaningful semantic organization. Internal metrics should be interpreted cautiously because they are influenced by cluster shape, density, and dimensionality and may not fully capture complex semantic structures; therefore, they are considered together with external metrics and qualitative analysis.

3.4 Thematic Structure of Clusters for Full-Text Representation

Documents are mapped to seven major academic fields: *Architecture and Spatial Planning* (F1), *Geology and Earth Sciences* (F2), *Electrical Systems and Energy* (F3), *Mechanical Systems and Industrial Processes* (F4), *Mathematical and Physics Sciences* (F5), *Construction and Infrastructure* (F6), and *Artificial Intelligence and Computer Science* (F7). Semantic clustering has been demonstrated to identify conceptual relationships beyond rigid disciplinary boundaries

[3]. Table 5 shows the distribution of clusters across disciplines.

Table 5. Distribution of documents by clusters (0-6) and academic fields

Cluster	F1	F2	F3	F4	F5	F6	F7	Total
0	21	0	0	0	2	6	0	29
1	3	1	3	0	51	9	3	70
2	0	0	12	0	0	1	0	13
3	2	0	5	5	3	4	0	19
4	1	4	7	2	2	7	6	29
5	0	18	2	2	2	1	0	25
6	3	0	0	0	9	31	2	45

Source: created by the authors

Clusters 0, 1, 2, 5, and 6 are more clearly defined by their dominant topics. Cluster 0 mainly includes documents from *Architecture and Spatial Planning*, Cluster 1 from *Mathematical and Physics Sciences*, Cluster 2 from *Electrical Systems and Energy*, Cluster 5 from *Geology and Earth Sciences*, and Cluster 6 from *Construction and Infrastructure*. *Mathematical and Physics Sciences* (F5) appear across several clusters, which reflects their broad role across different disciplines. A similar pattern is observed for *Artificial Intelligence and Computer Science* (F7), which appear in several semantic areas rather than forming a single cluster. Clusters 3 and 4 have a more mixed composition, combining documents from engineering, applied sciences, and computational fields. This suggests that they include interdisciplinary topics rather than one clearly dominant theme. The full-text representation therefore shows both thematic alignment with academic fields and overlaps between disciplines. These relationships are further examined through cosine similarity and inter-cluster comparisons.

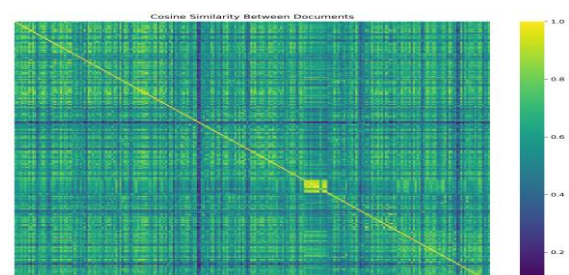


Fig. 4: Cosine similarity matrix

Source: created by the authors

The cosine similarity matrix shown in Figure 4 reveals heterogeneous similarity patterns across documents. The values along the diagonal represent each document comparing with itself, while higher values outside the diagonal indicate documents that are more similar in content. Moderate similarities

between clusters show that they are not completely isolated and share certain concepts, reflecting the overlap of topics in scientific texts. As shown in Figure 5, most cosine similarity values calculated from pairwise document comparisons fall within the range of 0.5-0.7.

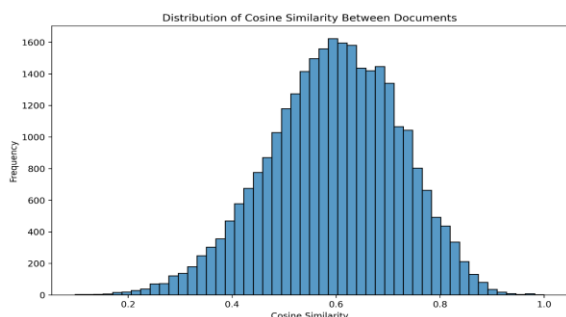


Fig. 5: Distribution of cosine similarity
Source: created by the authors

This indicates a balance between semantic cohesion and differentiation. Excessively high similarity would suggest cluster overlap, whereas exceptionally low similarity would hinder the formation of meaningful groups. The observed pattern therefore suggests a continuous but structured semantic space.

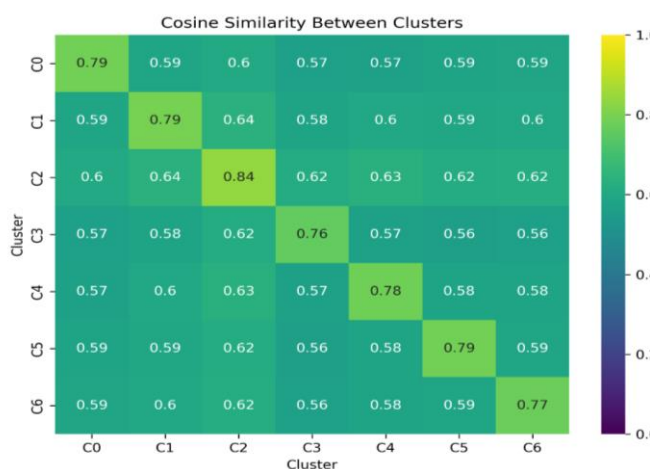


Fig. 6: Cluster-to-cluster similarity
Source: created by the authors

The cluster-to-cluster similarity map presented in Figure 6 is computed by using the average pairwise cosine similarity between documents that belong to different clusters. It provides a higher-level view of the semantic relationships between clusters and helps compare their content-based similarity. Within-cluster similarity values range from 0.76 to 0.84, indicating strong internal cohesion. In contrast, between-cluster similarity values range from 0.56 to 0.64, showing that the

clusters remain distinguishable while still sharing some common concepts.

This pattern indicates that the semantic space is locally coherent but remains connected at a broader level through shared thematic elements. This can be observed in both the visual representations and the similarity values, which are consistent with the clustering evaluation metrics and make the semantic structure easier to interpret.

3.5 Thematic Structure of Clusters for Section-Based Representation

The section-based representation improves semantic density and clustering performance. The results presented in Table 6 indicate clearer and more structured semantic organization compared with the full-text representation. Clusters show improved thematic homogeneity, with clearer boundaries and reduced semantic overlap. Most clusters are dominated by a single academic field, indicating improved thematic separation. Cluster 2 is almost entirely associated with Electrical Systems and Energy (F3), Cluster 5 is dominated by Geology and Earth Sciences (F2), and Cluster 6 corresponds mainly to Construction and Infrastructure (F6). Cluster 1 is largely associated with Mathematical and Physics Sciences (F5), reflecting a strong conceptual concentration within this field.

Table 6. Distribution of documents by clusters and academic fields (section-based)

Cluster	F1	F2	F3	F4	F5	F6	F7	Total
0	26	0	0	0	1	2	0	29
1	1	0	2	0	56	9	1	69
2	0	0	16	0	0	0	0	16
3	0	0	2	7	0	2	0	11
4	0	2	5	1	0	4	9	21
5	0	20	2	1	1	1	0	25
6	2	0	0	0	9	41	1	53

Source: created by the authors

Compared with the full-text representation, the section-based results are less fragmented and follow the disciplinary structure more closely. However, clusters 3 and 4 remain more heterogeneous, containing documents from mechanical systems, energy, and artificial intelligence. These clusters represent areas where concepts and methods from different disciplines come together. Using sections with more concentrated semantic information produces more homogeneous clusters that are better separated and easier to interpret.

3.6 Full-Text vs Section-Based Comparison

To evaluate the impact of representation strategy on the semantic space, the full-text and section-based representations are compared within the same analytical pipeline. The section-based representation improves cluster cohesion, separation, and most evaluation metrics compared with the full-text representation.

Table 7. Clustering performance metrics for full-text and section-based representations

Metric	Full-Text	Section-Based
Purity	0.63	0.647
NMI	0.3968	0.4806
ARI	0.35	0.3343
Silhouette	0.5349	0.7154
Davies-Bouldin	0.7608	0.5025
Calinski-Harabasz	219.37	651.99

Source: created by the authors

Table 7 shows that the section-based representation improves most clustering performance indicators. Internal metrics show a substantial increase in cohesion and separation, with a higher Silhouette score, a lower Davies-Bouldin index, and a considerably higher Calinski-Harabasz value [21], [22], [23]. External metrics also improve, particularly NMI, which increases from 0.3968 to 0.4806. In contrast, ARI decreases slightly from 0.3500 to 0.3343, indicating that improved cluster compactness does not necessarily result in stronger exact agreement with the predefined disciplinary labels [24].

To statistically assess the association between clusters and academic fields, the chi-square test of independence and Cramer's V are computed for both representations. The full-text representation obtains $\chi^2 = 447.778$, $p = 1.58 \times 10^{-72}$, and Cramer's $V = 0.570$, while the section-based representation obtains $\chi^2 = 748.443$, $p = 4.90 \times 10^{-134}$, and Cramer's $V = 0.746$. The higher Cramer's V value for the section-based representation indicates a stronger association between the resulting clusters and academic fields. This provides statistical evidence of a clearer, more compact, and more discipline-aligned semantic structure under the section-based representation.

In the full-text representation shown in Figure 7, clusters appear more dispersed and exhibit greater overlap. In contrast, the section-based representation shown in Figure 8 produces more compact and better-separated clusters with clearer boundaries.

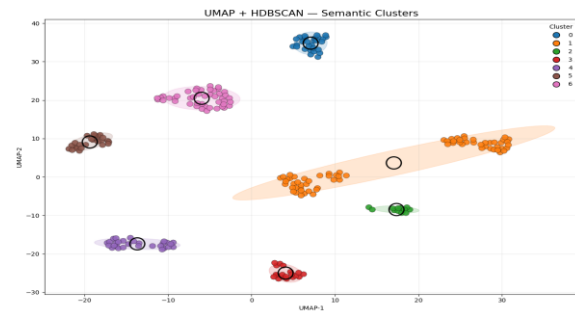


Fig. 7: UMAP visualization (full-text)

Source: created by the authors

The agreement between the quantitative metrics and the visualizations supports the interpretation that the section-based representation produces a more compact and clearly separated semantic structure.

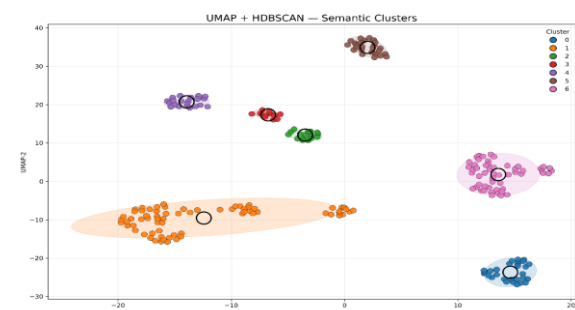


Fig. 8: UMAP visualization (section-based)

Source: created by the authors

The section-based representation produces a more selective density-based structure as illustrated in Figure 8, in which semantically ambiguous or boundary documents may receive the noise label rather than be assigned to clusters with weak local density support. In this context, noise assignments provide a practical indicator of clustering uncertainty. The observed improvement is related to semantic density: abstracts, introductions, and conclusions contain more concentrated information about research objectives, methods, and findings, while reducing the local semantic variation represented by ϵ_{ij} in Equation (2). As a result, the section-based representation improves cohesion, separation, and interpretability, confirming the importance of representation choice in unsupervised semantic clustering.

Transformer-based embeddings combined with UMAP and HDBSCAN provide a robust framework for uncovering latent thematic structures in large-scale Albanian scientific corpora, consistent with previous studies on semantic and text clustering [25], [26]. The results show that the representation strategy strongly affects clustering performance: the section-based representation produces more

compact, well-separated, and interpretable clusters than the full-text representation, in line with prior work on text representation [25], [27]. BERTopic, spectral clustering, graph-based methods, and contrastive representation learning remain relevant alternatives for semantic clustering and topic discovery [3], [9], [25], [28]. The SBERT-UMAP-HDBSCAN framework is retained due to its unsupervised, scalable, and interpretable design, making it suitable for long-form Albanian scientific texts with limited annotations. However, it remains sensitive to UMAP hyperparameters and segmentation choices [15]. The section-based representation may omit contextual details present in full-text data, and the findings may require adaptation for other languages or domains [25]. This work provides new insight into representation strategies for semantic clustering in Albanian and low-resource language settings [4], [5], [6], [7], [8]. Future work will include systematic hyperparameter sensitivity analysis and comparisons with additional clustering frameworks such as BERTopic and graph-based methods.

4 Conclusions

This study presents a comprehensive analysis of the semantic structure of a large-scale corpus of Albanian scientific texts using transformer-based embeddings and unsupervised clustering techniques. The results demonstrate that the proposed framework effectively captures meaningful thematic structures in the absence of predefined labels, revealing both well-defined domain-specific clusters and interdisciplinary boundary regions. A key contribution of this work is the demonstration that text representation plays a critical role in determining clustering performance. The section-based representation improves cluster cohesion and separation compared with the full-text representation, resulting in a more structured and interpretable semantic space. The findings further suggest that increasing embedding dimensionality does not necessarily improve clustering performance in heterogeneous scientific corpora. Interpreting noise-assigned documents as semantically ambiguous or boundary cases provides additional insight into the continuous nature of the semantic space. Overall, the study highlights the importance of representation design in semantic modeling, particularly in low-resource language settings. The proposed framework offers a practical and scalable approach for analyzing large scientific corpora and may be adapted to other languages and low-resource research domains.

Future work will focus on hybrid representation strategies, adaptive segmentation methods, and broader comparisons with advanced clustering and topic modeling approaches, including BERTopic variants, spectral clustering, graph-based representations, and contrastive embedding models.

References:

- [1] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 3982-3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>.
- [2] Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2025). Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6, 100-108. <https://doi.org/10.1016/j.ijcce.2024.11.004>.
- [3] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. <https://doi.org/10.48550/arXiv.2203.05794>.
- [4] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282-6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>.
- [5] Hoxha, K., & Baxhaku, A. (2018). An automatically generated annotated corpus for Albanian named entity recognition. *Cybernetics and Information Technologies*, 18(1), 95-108. <https://doi.org/10.2478/cait-2018-0009>.
- [6] Mati, D. N., Hamiti, M., & Mollakuqe, E. (2021). Morphological tagging and lemmatization in the Albanian language. *SEEU Review*, 16(2), 3-16. <https://doi.org/10.2478/seeur-2021-0015>.
- [7] Çano, E., & Lamaj, D. (2024). AlbNews: A corpus of headlines for topic modeling in Albanian. arXiv. <https://arxiv.org/abs/2402.04028>.
- [8] Kaçorri, D., Basholli, A., & Prifti, L. (2024). Using cluster analysis for author classification of Albanian texts: A study on the

effectiveness of stop words. WSEAS Transactions on Computer Research, 12, 19-28.

<https://doi.org/10.37394/232018.2024.12.2>.

- [9] Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6894-6910). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2021.emnlp-main.552>.
- [10] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. Journal of Open Source Software, 2(11), 205.
<https://doi.org/10.21105/joss.00205>.
- [11] Ortakci, Y., & Borhan, B. (2025). Optimizing SBERT for long text clustering: Two novel approaches with empirical insights. *The Journal of Supercomputing*, 81(8), Article 950. <https://doi.org/10.1007/s11227-025-07414-4>.
- [12] Khan, T. F., Arslan, M., Alvi, M., Khan, M. A., & Mehmood, A. (2025). LLM-based enhanced clustering for low-resource language: An empirical study. *Computer Modeling in Engineering & Sciences*. <https://doi.org/10.32604/cmescs.2025.068887>.
- [13] Curma, H., & Sinaj, V. (2026). Automating the classification of economic activities in official statistics: A comparative study of neural networks and transformers. WSEAS Transactions on Computer Research, 14.
<https://doi.org/10.37394/232018.2026.14.4>.
- [14] Falconí, D., & Marrero, S. (2025). Methods to determine the similarity and distance between researchers from classification algorithms. WSEAS Transactions on Computer Research, 13, 138-145.
<https://doi.org/10.37394/232018.2025.13.14>.
- [15] McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. <https://arxiv.org/abs/1802.03426>.
- [16] Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science, Vol. 7819, pp. 160-172). Springer.
https://doi.org/10.1007/978-3-642-37456-2_14.
- [17] Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), [Online]. <https://openreview.net/forum?id=SyK00v5xx> (Accessed Date: January 22, 2026).
- [18] Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review, 53(2), 217-288.
<https://doi.org/10.1137/090771806>.
- [19] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605, [Online].
<https://www.jmlr.org/papers/v9/vandermaten08a.html> (Accessed Date: January 5, 2026).
- [20] Abdallah, L., Hijazi, M., & Badarna, M. M. (2026). K-means clustering algorithm for mixed data using multidimensional scaling. WSEAS Transactions on Computer Research, 14, 163-169.
<https://doi.org/10.37394/232018.2026.14.14>.
- [21] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [22] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224-227.
<https://doi.org/10.1109/TPAMI.1979.4766909>.
- [23] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1-27.
<https://doi.org/10.1080/03610927408827101>.
- [24] Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193-218. <https://doi.org/10.1007/BF01908075>.
- [25] Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In C. C. Aggarwal & C. Zhai (Eds.), Mining Text Data (pp. 77-128). Springer.
https://doi.org/10.1007/978-1-4614-3223-4_4.
- [26] Nazemi, K., Burkhardt, D., Ginters, E., & Kohlhammer, J. (2015). Semantics visualization: Definition, approaches and challenges. Procedia Computer Science, 75, 75-83.
<https://doi.org/10.1016/j.procs.2015.12.216>.

- [27] Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43-52. <https://doi.org/10.1007/s13042-010-0001-0> (Accessed Date: January 5, 2026).
- [28] Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. Technical Report No. 00-034, Department of Computer Science and Engineering, University of Minnesota, [Online]. <https://conservancy.umn.edu/handle/11299/215421> (Accessed Date: January 5, 2026).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Leona Rexhaj: Conceptualization, methodology, data curation, software, writing - original draft, writing - review & editing. Luella Prifti: Conceptualization, methodology, writing - review & editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The Polytechnic University of Tirana financially supports the publication of this article.

Conflict of Interest

The authors declare that they have no conflict of interest regarding the publication of this paper.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0

https://creativecommons.org/licenses/by/4.0/deed.en_US