A Per-Cluster Multi-Scale Topic Modeling Framework with Unstructured and Heterogeneous Text Corpora

Mamadou Alpha Hawa BALDE[1,2,3][0009−0000−6096−2082], Jassim BENSAFIR[2,3], Sarah BEN OTHMAN[1,2], Salah ZIDI[3,4], and Slim HAMMADI[1,2]

[1] CRIStAL UMR 9189, Avenue Henri Poincaré, 59655 Villeneuve d'Ascq, FRANCE
https://www.cristal.univ-lille.fr/
[2] Centrale Lille Institut, Cité Scientifique, 59650 Villeneuve-d'Ascq, FRANCE
https://centralelille.fr/
[3] NeoLedge, 49 Bd de Strasbourg, 59042 Lille Cedex, FRANCE
https://www.neoledge.com/
[4] IResCoMath Lab,University of Gabes, Gabes, TUNISIA
https://irescomath-lab.org/

Abstract. Local government systems produce large, heterogeneous text corpora, including policy reports, internal communications, and public documentation. Extracting interpretable topics from such corpora of unstructured data requires methods that capture both global themes and fine-grained, domain-specific subtopics. Transformer-based topic modeling frameworks such as BERTopic provide effective embeddings and clustering, but their reliance on a single global clustering step imposes uniform topic granularity, often merging semantically distinct discourse regions. We introduce Per-Cluster Topic Modeling (PCTM), a scalable extension to BERTopic that performs neighbor-aware local clustering of document embeddings followed by per-cluster topic extraction. This approach generates multi-scale topic representations, enabling coarse-grained administrative themes and fine-grained subtopics to coexist within the same framework. Evaluated on a municipal office corpus spanning multiple departments, PCTM outperforms global BERTopic baselines in topic coherence, stability, and interpretability. Beyond empirical gains, PCTM provides a computational model of multi-level social systems, reflecting the heterogeneous semantic structure of municipal governance. This framework supports AI-driven analysis of complex administrative and social text corpora, offering a robust tool for structured knowledge discovery.

Keywords: BERTopic · Embedding-Based Topic Modeling · Multi-Scale Topic Representation · Local Neighbor-Aware Clustering · Heterogeneous Text Corpora · Social Systems Modeling.

1 Introduction

In a local municipal administration the amount of text data generated as part of their routine governance processes are tremendous [16]. These data include

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

policy reports, internal memoranda, public communications, and administrative documentation. All together, these texts form complex and heterogeneous corpora that reflect the structure and functioning of municipal governance as a social system. Different administrative units, policy domains, and communicative contexts operate at distinct semantic levels, ranging from high-level strategic discourse to highly specific operational detail [23]. Effectively modeling such corpora therefore requires computational methods capable of representing both global thematic structure and fine-grained, domain-specific variation.

Topic modeling has long served as a foundational approach for extracting latent thematic structure from large text collections [21, 12, 17]. Probabilistic models such as Latent Dirichlet Allocation (LDA) and its hierarchical extensions provide interpretable topic representations but rely on bag-of-words assumptions that limit their effectiveness on short, noisy, or semantically diverse documents [9, 4]. More recently, embedding-based topic modeling approaches have leveraged transformer language models to capture contextual semantic information, enabling improved performance on heterogeneous text corpora [20, 18, 7]. Among these, BERTopic has emerged as a widely adopted framework that combines document embeddings, density-based clustering, and class-based TF-IDF to produce coherent and interpretable topics [6].

Despite its effectiveness, BERTopic imposes a critical modeling assumption that is problematic in socially heterogeneous settings. This is the formation of a single global clustering step which is applied across the entire embedding space . As a result, it implicitly enforces a uniform topic granularity for all documents. In municipal corpora, however, semantically distinct discourse regions often coexist, corresponding to different administrative subsystems with inherently different levels of abstraction. Applying a global clustering resolution in such contexts can merge locally coherent but globally distant discourse, leading to diluted topics, unstable cluster assignments, and reduced interpretability. This limitation becomes particularly noticeable when analyzing municipal governance as a multi-level social system.

From a computational modeling perspective, municipal text corpora can be viewed as composed of multiple interacting subsystems, each characterized by its own semantic structure and scale [23]. Modeling such systems requires adaptive representations that respect local coherence while maintaining a global view of the system. Fixed-scale topic modeling approaches, whether probabilistic or embedding-based, are not suitable to this task, as they assume homogeneity in thematic resolution across the corpus. This motivates the need for topic modeling frameworks that can operate at multiple semantic scales in a data-driven and locally adaptive manner.

In this work, we propose Per-Cluster Topic Modeling (PCTM), a scalable extension of BERTopic designed to address the heterogeneity of municipal text corpora. Rather than applying topic modeling to the entire corpus at a single resolution, PCTM first decomposes the document embedding space into locally coherent clusters using neighbor-aware clustering. Topic modeling is then performed independently within each cluster, allowing topic granularity to adapt

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

to the semantic characteristics of each discourse region. This per-cluster approach yields a multi-scale topic representation that captures both high-level administrative themes and fine-grained, subject-specific discourse within a unified computational framework.

We evaluate PCTM on a real-world administrative office dataset spanning multiple administrative domains and communicative functions. Through quantitative metrics and qualitative analysis, we demonstrate that PCTM improves topic coherence, stability, and interpretability compared to standard BERTopic. Beyond empirical performance, the proposed framework provides a computational abstraction that more faithfully reflects the multi-level structure of municipal governance, supporting the analysis of municipal communication as a complex social system.

The main contributions of this paper are summarized as follows:

1. We identify limitations of global-scale embedding-based topic modeling when applied to heterogeneous municipal text corpora.
2. We introduce PCTM, a per-cluster multi-scale topic modeling framework that adaptively selects topic granularity based on local semantic structure.
3. We empirically demonstrate the effectiveness of the proposed approach on real-world municipal data and discuss its implications for AI-driven social system modeling.

The remainder of this paper is structured as follows. Section 2 presents the background and reviews related work relevant to this study. Section 3 describes the proposed methodology in detail. The experimental setup and results are reported in Section 4, followed by a discussion of the findings in Section 5. Finally, Section 6 concludes the paper by summarizing the main contributions and outlining directions for future work.

2 Background and Related Work

2.1 Topic Modeling for Heterogeneous Text Corpora

Topic modeling has been widely used to uncover latent thematic structure in large text corpora [21, 2, 11]. Probabilistic approaches such as Latent Dirichlet Allocation (LDA) and its extensions provide interpretable topic representations but rely on bag-of-words assumptions that limit their effectiveness for short texts and semantically heterogeneous corpora [9, 8, 13]. Hierarchical variants, including hierarchical LDA and nonparametric models such as the Hierarchical Dirichlet Process, introduce topic hierarchies but require predefined structural assumptions and often struggle to scale or adapt to diverse real-world datasets.

In socially complex domains such as governance and public administration, text corpora are often heterogeneous not only in content but also in semantic granularity [20, 14, 10, 19]. Documents may range from high-level policy statements to highly specific operational records, challenging models that assume uniform thematic resolution [3, 22]. This has motivated increasing interest in topic modeling approaches that can accommodate heterogeneity and multi-level structure [23, 22].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

2.2 Embedding-Based Topic Modeling

Recent advances in transformer-based language models have enabled embedding-based topic modeling methods that capture contextual semantic information beyond word co-occurrence statistics [6, 7, 20, 5]. Approaches such as Top2Vec and BERTopic cluster document embeddings and derive topic representations from clustered document sets, improving performance on short and noisy texts [1, 7, 20]. Among these, BERTopic has gained prominence due to its modular design, combining transformer embeddings, density-based clustering, and class-based TF-IDF to produce interpretable topics [6].

Despite these advantages, embedding-based topic models typically apply a single global clustering step, implicitly enforcing a uniform topic granularity across the corpus. While effective for relatively homogeneous datasets, this assumption becomes problematic for heterogeneous corpora, where semantically distinct discourse regions may require different levels of abstraction. As a result, global clustering can merge locally coherent topics or produce unstable topic representations in complex social datasets.

2.3 Multi-Scale and Hierarchical Topic Representations

Multi-scale and hierarchical topic modeling approaches aim to capture thematic structure at different levels of abstraction. Traditional probabilistic models achieve this through explicit hierarchical priors, while embedding-based frameworks often rely on post-hoc topic reduction or hierarchical clustering of topic representations. In BERTopic, for example, hierarchical topic modeling and topic merging can be applied after initial topic extraction to produce coarser views of the topic space [6].

However, these approaches typically operate after global topic extraction and do not adapt topic granularity during the clustering process itself. Consequently, they may not fully address local heterogeneity in the embedding space, particularly in corpora where different discourse regions exhibit fundamentally different semantic scales. This limits their effectiveness for modeling socially complex systems such as municipal governance.

2.4 Topic Modeling for Social and Administrative Systems

Topic modeling has been applied extensively to social and administrative text data, including policy documents, government reports, and public communications [23, 3, 16]. These studies demonstrate the value of topic models for understanding institutional priorities, policy evolution, and public discourse [15]. However, many applications rely on fixed-scale topic models and focus primarily on empirical insights rather than modeling assumptions.

From a computational modeling perspective, municipal governance can be viewed as a multi-level social system composed of interacting subsystems, each characterized by distinct communicative practices. Modeling such systems requires representations that reflect both global coordination and local specialization. These interactions are reflected in the text through shared vocabulary and

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

co-occurring themes across documents. However, topic models such as PCTM do not explicitly represent these relationships, instead capturing them indirectly through patterns in word use and topic co-occurrence, where interactions emerge as latent statistical structure. This perspective motivates adaptive topic modeling approaches that align more closely with the structural properties of social systems.

2.5 Positioning of the Present Work

In contrast to prior approaches, the proposed Per-Cluster Topic Modeling (PCTM) framework introduces adaptive topic granularity at the cluster level, enabling multi-scale representations that better capture semantic heterogeneity in municipal corpora. By moving beyond global clustering strategies, PCTM provides a more flexible alternative for diverse textual data, addressing limitations of globally uniform topic extraction.

3 Methodology

This section presents the PCTM framework, a scalable extension of BERTopic. The approach partitions the embedding space into locally coherent clusters, after which topic modeling is performed independently within each cluster. This design allows topic extraction to adapt to local semantic structure, reflecting variations in semantic density across the data.

3.1 Per-Cluster Topic Modeling (PCTM)

Problem Definition Let $D = \{d_1, d_2, \ldots, d_N\}$ represent a corpus of N municipal documents, where each document may originate from distinct administrative units, policy domains, or communicative contexts.

The objective is to learn a set of topics $T = \{T_1, T_2, \ldots, T_K\}$ that captures semantic structure at multiple scales, preserving local coherence while representing global thematic trends across heterogeneous corpora.

Document Embeddings : Each document d_i is embedded into a semantic vector $e_i \in \mathbb{R}^m$ using a pre-trained transformer model f_{emb} :

$$e_i = f_{emb}(d_i).$$

The embeddings of all documents form the matrix

$$E = [e_1, e_2, \ldots, e_N].$$

Local Neighbor-Aware Clustering Standard BERTopic applies a single global clustering step over E, enforcing uniform topic granularity. PCTM addresses this limitation by partitioning E into locally coherent clusters, enabling per-cluster granularity adaptation. This clustering is performed based on the neighborhood graph contruction and subsequent cluster detection.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

Neighborhood Graph Construction A k-nearest neighbor (k-NN) graph $G = (V, E)$ is constructed from the document embeddings, where V is the set of nodes (each representing a document) and E is the set of edges connecting each document to its k nearest neighbors based on cosine similarity:

$$w_{ij} = \cos(e_i, e_j).$$

Cluster Detection Using density-based clustering, the graph G is partitioned into C clusters with HDBSCAN:

$$\mathcal{C} = \{C_1, C_2, \ldots, C_C\}, \quad C_c \subseteq D.$$

Where each cluster C_c contains a subset of the documents D. Each cluster corresponds to a semantically coherent subspace of the document embedding space.

Per-Cluster Topic Extraction For each cluster C_c, topic modeling is performed independently:

$$\mathcal{T}_c = \{T_{c,1}, T_{c,2}, \ldots, T_{c,K_c}\}.$$

The number of topics K_c is adaptive, allowing clusters with broad semantic coverage to produce fewer coarse-grained topics, while narrower clusters yield more fine-grained topics.

Cluster-specific topics are derived using class-based TF–IDF (c-TF–IDF):

$$\text{cTFIDF}_{c,j} = \frac{f_{c,j}}{\sum_{t \in V_c} f_{c,t}} \cdot \log\left(\frac{N}{n_j}\right),$$

where $f_{c,j}$ denotes the frequency of term j in cluster C_c, $\sum_{t \in V_c} f_{c,t}$ is the total number of term occurrences in C_c, n_j is the number of clusters containing term j, and N is the total number of clusters. This weighting emphasizes terms that are frequent within a cluster but rare across clusters. Topics are then constructed by selecting the top-ranked terms according to c-TF-IDF within each cluster, with each topic represented as a set of its highest scoring terms. For example, consider a cluster related to urban planning where terms such as zoning, permit, and construction appear frequently. The c-TF-IDF weighting increases the importance of these terms if they are common within this cluster but relatively rare across other clusters. This results in topic representations that are both locally representative and globally discriminative.

Multi-Scale Topic Representation By independently modeling topics per cluster, PCTM produces a multi-scale topic representation:

- Coarse-grained topics for broad clusters
- Fine-grained topics for semantically tight clusters

This allows the model to reflect the multi-level structure of municipal governance, where different administrative or policy domains operate at distinct semantic scales.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

Algorithm Summary: Per-Cluster Topic Modeling (PCTM)

Algorithm: Per-Cluster Topic Modeling (PCTM)
Input: Document corpus D, embedding model f_{emb}, neighborhood size k
Output: Multi-scale topic set $\mathcal{T}$

1. Compute document embeddings $E = \{f_{emb}(d) \mid d \in D\}$
2. Construct a k-NN graph G from E
3. Partition G into locally coherent clusters $\mathcal{C} = \{C_1, \ldots, C_{|\mathcal{C}|}\}$ using HDBSCAN clustering method
4. For each cluster $C_c \in \mathcal{C}$:
   (a) Apply BERTopic topic extraction restricted to documents in C_c
   (b) Extract cluster-specific topics $\mathcal{T}_c$
5. Return $\mathcal{T} = \bigcup_c \mathcal{T}_c$

Note that the initial clustering defines semantic regions rather than topics. Topic granularity is learned independently within each region.

Computational Properties

- **Scalability:** Clustering and topic extraction are performed on smaller subspaces rather than the full corpus, and are executed independently across clusters, enabling efficient and parallel processing of large text corpora.
- **Stability:** Local clustering mitigates sensitivity to global embedding distortions.
- **Interpretability:** Topics align with administrative or policy subsystems, providing interpretable representations of municipal governance.

Architecture of the proposed Per-Cluster Topic Modeling (PCTM) framework is illustrated on figure 1.

Evaluation Metrics We evaluate PCTM topics using three complementary metrics:

Topic coherence measures the semantic consistency of top words in each topic, computed per cluster and averaged across clusters to capture both local and global structure:

$$C_{PCTM} = \frac{1}{C}\sum_{c=1}^{C} C(C_c) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{K_c}\sum_{k=1}^{K_c} C_{UMass}(T_{c,k}) \tag{1}$$

where C is the number of clusters, K_c the number of topics in cluster C_c, $T_{c,k}$ denotes the k-th topic of cluster c, and $C_{UMass}()$ is the UMass topic coherence score. The UMass coherence metric evaluates topic quality based on word co-occurrence statistics in the corpus. It measures how frequently the top words of a topic appear together within documents, with higher scores indicating greater semantic consistency.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

Topic stability quantifies reproducibility via the average Jaccard similarity of top-n words across multiple runs, aggregated per cluster and overall:

$$S_{PCTM} = \frac{1}{C}\sum_{c=1}^{C} S(C_c) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{K_c}\sum_{k=1}^{K_c} S(T_{c,k}) \tag{2}$$

where $S(T_{c,k})$ denotes the stability of topic k in cluster c, computed as the average Jaccard similarity of its top-n words across runs.

Topic diversity measures how distinct topics are within each cluster via the proportion of unique words among all top-n words, averaged across clusters:

$$D_{PCTM} = \frac{1}{C}\sum_{c=1}^{C} D(C_c) \tag{3}$$

where $D(C_c)$ denotes the topic diversity within cluster C_c, computed as the proportion of unique words among the top-n words of all topics in the cluster.

Together, these metrics provide a comprehensive assessment of the quality, robustness, and distinctiveness of the discovered topics.

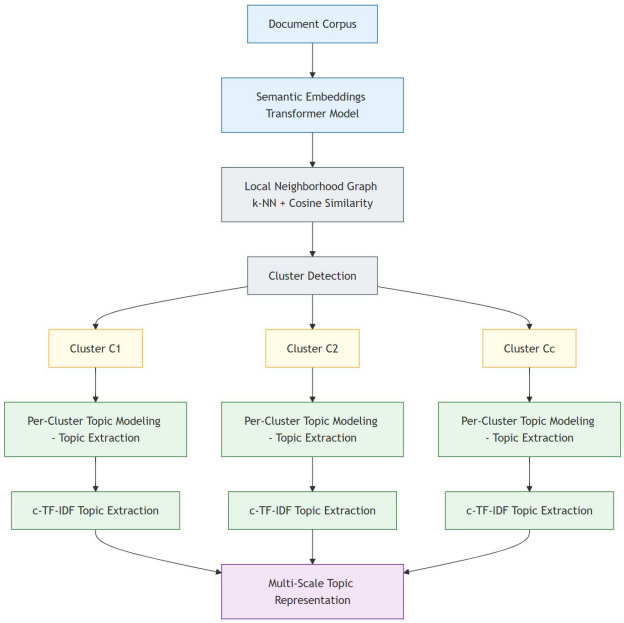

Fig. 1. Architecture of the proposed PCTM framework
Documents are embedded into a semantic space and organized into a local neighborhood graph. The embedding space is partitioned into semantically coherent clusters, within which topic modeling is performed independently using adaptive topic counts and c-TF-IDF. This yields a multi-scale topic representation that captures both global and local semantic structure.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

Social System Perspective From a social systems modeling viewpoint, clusters correspond to functional subsystems within municipal governance, each requiring an adaptive level of semantic resolution. PCTM thus provides a computationally grounded abstraction that mirrors the multi-level structure of municipal social systems.

4 Results

In this section, We evaluate the proposed PCTM framework on a real-world administrative corpus, comparing it to standard BERTopic via quantitative metrics and qualitative analysis to test its ability to capture the heterogeneous, multi-level structure of municipal administrative communication.

4.1 Dataset

We analyze 10,000 municipal correspondence records from a French local authority, including reports, announcements, and internal communications that reflect diverse operational contexts. The dataset is confidential due to privacy constraints, with all personal or sensitive information removed prior to processing.

Each record includes:

- A **subject line** (mandatory)
- Optional **content or attachment text**
- Administrative metadata:
  - **Mail Type** (urban planning, personnel management)
  - **Sender Type** (administration, population, enterprise, association)
- Average document length: 220 words
- Preprocessing: lowercasing, punctuation removal, lemmatization
- Embedding model: S-BERT (paraphrase-multilingual-MiniLM-L12-v2)

The corpus covers a wide range of administrative functions, communicative intents, and actors, resulting in strong semantic heterogeneity and variable granularity. This reflects real-world municipal communication and challenges fixed-resolution topic models.

4.2 Baselines

- **BERTopic (global):** Standard BERTopic applied to the full corpus using a single global clustering step, enforcing uniform topic granularity.

PCTM differs by first partitioning the embedding space into locally coherent clusters using density-based clustering, and then applying BERTopic independently within each cluster, allowing the number and granularity of topics to adapt to local semantic structure.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

4.3 Quantitative Results

The dataset was divided into 33,526 chunks, each embedded as a 384-dimensional vector. PCTM identified 127 clusters such as Cluster 78 with 202 chunks. In total, 139 topics were discovered and the number of topics per cluster ranged from 1 to 12.

Table 1. Comparison of topic modeling methods across coherence, stability, and diversity metrics. Shaded cells here indicate the best performance.

Method	Coherence (c_v)	Stability	Diversity
BERTopic (Global)	0.41	0.62	0.45
PCTM (Ours)	**0.58**	**0.79**	**0.52**

Topic Coherence We calculate topic coherence per cluster and report the overall PCTM coherence by averaging across clusters, ensuring that both local and global semantic structures are captured.
Observation: As shown in Table 1, PCTM achieves an average topic coherence of 0.58, compared to 0.41 for BERTopic. This demonstrates that PCTM produces more semantically consistent topics and better alignment within locally coherent clusters.

Topic Stability We computed topic stability by measuring the average Jaccard similarity of the top-n words across multiple runs with different embedding seeds. Stability is first averaged across topics within each cluster and then across all clusters, providing an overall measure of how consistently the model captures semantic patterns.
Observation: PCTM demonstrates substantially higher stability than global BERTopic. By restricting topic extraction to local embedding neighborhoods, PCTM reduces sensitivity to global embedding perturbations and clustering variability, leading to more robust topic representations.

Topic diversity Topic diversity measures how distinct the topics are by computing the proportion of unique words across all top-n words in each cluster. The cluster-level diversity is averaged across all clusters to yield an overall PCTM diversity score. Higher values indicate more semantically distinct topics.

Multi-Scale Topic Distribution PCTM generates topics at multiple semantic scales:

- Coarse-grained clusters: high-level administrative themes (example: *Urban Planning Strategy*)
- Fine-grained clusters: department-specific subtopics (example: *Public Safety: Traffic Enforcement Policies*)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

Global BERTopic fails to capture these distinctions, as it imposes a uniform scale.

In terms of computational cost, PCTM introduces an additional clustering step and multiple local topic modeling runs compared to standard BERTopic. Although this adds overhead, operating on smaller data subsets reduces memory usage and enables parallel execution. In practice, runtime remains comparable or moderately higher depending on the number of clusters while significantly improving topic quality.

4.4 Qualitative Analysis

Qualitative inspection of extracted topics highlights PCTM's benefits. Clusters capture major domains like urban planning or citizen services while within each, finer topics reflect operational subthemes such as permits, procedures, or complaints.

In contrast, global BERTopic often merges distinct administrative functions into single topics or fragments coherent domains across multiple topics, reflecting the limitations of fixed-scale global clustering.

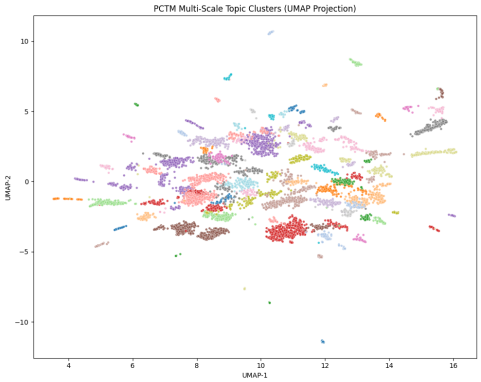

Fig. 2. UMAP projection of document embeddings colored by PCTM clusters.

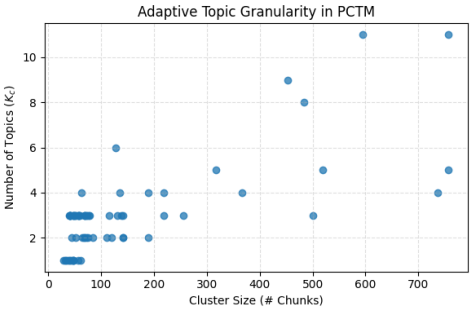

Fig. 3. Relation between cluster size and number of topics.

Figure 2 shows a two-dimensional UMAP projection of document embeddings colored by PCTM clusters. Clusters vary in size, density, and shape: compact clusters represent focused topics, while larger ones reflect broader themes. This variability supports per-cluster topic modeling, as a single global resolution cannot capture such heterogeneous structure.

Figure 3 illustrates the relationship between cluster size and the number of topics extracted per cluster in PCTM. The results show non-linear scaling: smaller clusters produce only a few detailed topics, while larger clusters generate more diverse topic sets. This adaptive behavior highlights the value of per-cluster topic modeling for heterogeneous corpora.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

4.5 Key Takeaways

1. PCTM outperforms global BERTopic in terms of coherence and stability on heterogeneous municipal correspondence.
2. Per-cluster topic extraction is the key causal mechanism, not clustering alone.
3. Per-cluster topic extraction reveals multi-scale semantic structure aligned with municipal social systems.
4. The framework is scalable, interpretable, and computationally grounded, making it suitable for AI-driven analysis of heterogeneous text corpora in administrative and social contexts.

5 Discussion

The experimental results demonstrate that Per-Cluster Topic Modeling (PCTM) provides substantial improvements over traditional global topic modeling approaches in both quantitative metrics and qualitative interpretability. By decomposing the embedding space into locally coherent clusters, PCTM allows adaptive topic granularity, which aligns with the multi-level structure of municipal governance. Although evaluated on municipal corpora, PCTM is applicable to any heterogeneous text collection with multi-scale semantic structure, such as healthcare records, legal documents, scientific literature, or customer feedback datasets. In such domains, we expect PCTM to outperform global topic models when semantic granularity varies across subdomains.

5.1 Implications for Social Systems Modeling

Municipal governance can be conceptualized as a complex social system, composed of interacting administrative units, policy domains, and communicative subsystems. PCTM captures this structure computationally:

- Clusters as subsystems: Each cluster corresponds to a semantically coherent administrative or policy domain.
- Multi-scale topics: Coarse-grained topics reflect strategic administrative themes, while fine-grained topics represent operational subdomains.
- Adaptive modeling: The method does not impose a uniform topic resolution, allowing representation of heterogeneous subsystems faithfully.

This capability is critical for AI-driven social system analysis, where fixed-resolution models often obscure local specialization and misrepresent subsystem interactions.

5.2 Methodological Contributions

From a computational perspective, PCTM introduces several innovations:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

- Local neighbor-aware clustering: Enhances topic stability by respecting embedding-space locality.
- Per-cluster topic extraction: Enables adaptive granularity, improving interpretability across heterogeneous corpora.
- Multi-scale representation: Integrates local and global semantic patterns without requiring hierarchical priors or post-hoc aggregation.

These design choices provide a scalable, interpretable, and reproducible framework suitable for municipal or other socially complex datasets.

5.3 Limitations and Future Work

While PCTM addresses key challenges in heterogeneous topic modeling, several limitations remain:

- Cluster quality dependency: Effectiveness depends on the quality of embeddings and local clustering; noisy or sparse embeddings may reduce performance.
- Implicit hierarchy: The multi-scale structure emerges from cluster density rather than an explicit hierarchy, which may limit some types of semantic analysis.
- Single-domain evaluation: Current experiments focus on a municipal corpus; broader generalization across other social systems (example: *healthcare, education*) remains to be tested.

Future work could explore explicit hierarchical extensions, cross-city datasets, or integration with temporal dynamics to model evolving social systems over time.

6 Conclusion

In this paper, we introduced Per-Cluster Topic Modeling (PCTM), a multi-scale extension of BERTopic designed for heterogeneous municipal text corpora. PCTM decomposes the embedding space into locally coherent clusters and performs per-cluster topic extraction, producing topics at variable semantic scales. Our evaluation demonstrates that PCTM:

1. Improves topic coherence and stability compared to global and fixed multi-scale BERTopic.
2. Produces multi-scale topics that align with administrative subdomains, reflecting the multi-level structure of municipal governance.
3. Provides a computationally interpretable framework suitable for AI-driven social system analysis.

By explicitly modeling semantic heterogeneity, PCTM offers a robust approach for analyzing complex social systems from textual data. This framework can support applications in policy analysis, administrative decision-making, and

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

public discourse modeling, while providing a foundation for future research in adaptive, multi-scale topic modeling across domains.

Overall, PCTM bridges the gap between embedding-based topic modeling and computational social system modeling, demonstrating how multi-scale, locally adaptive approaches can enhance both interpretability and analytical rigor in heterogeneous text corpora.

Acknowledgments. This study was funded by NeoLedge.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cong, L.W., Liang, T., Zhang, X., Zhu, W.: Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information (Nov 2024). https://doi.org/10.3386/w33168
2. Das, P., Mandal, S., Nedungadi, P., Raman, R.: Unveiling sustainable tourism themes with machine learning based topic modeling. Discover Sustainability **6**(1), 280 (Apr 2025). https://doi.org/10.1007/s43621-025-01065-4
3. David, A., Yigitcanlar, T., Desouza, K., Li, R.Y.M., Cheong, P.H., Mehmood, R., Corchado, J.: Understanding local government responsible AI strategy: An international municipal policy document analysis. Cities **155**, 105502 (Dec 2024). https://doi.org/10.1016/j.cities.2024.105502
4. Farea, A., Tripathi, S., Glazko, G., Emmert-Streib, F.: Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research. Engineering Applications of Artificial Intelligence **136**, 108877 (Oct 2024). https://doi.org/10.1016/j.engappai.2024.108877
5. Gokcimen, T., Das, B.: Topic Modelling Using BERTopic for Robust Spam Detection. In: 2024 12th International Symposium on Digital Forensics and Security (ISDFS). pp. 1–5 (Apr 2024). https://doi.org/10.1109/ISDFS60797.2024.10527342, iSSN: 2768-1831
6. Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure (Mar 2022). https://doi.org/10.48550/arXiv.2203.05794, arXiv:2203.05794 [cs]
7. Islam, K.M.S., Karri, R.T., Vegesna, S., Wu, J., Madiraju, P.: Contextual Embedding-based Clustering to Identify Topics for Healthcare Service Improvement. In: 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC). pp. 794–799 (Jul 2025). https://doi.org/10.1109/COMPSAC65507.2025.00106, iSSN: 2836-3795
8. Jiang, S., Li, H., Gan, D.: Technology acceptance model for online education: identifying interdisciplinary topics and their evolution based on BERTopic model. Social Sciences & Humanities Open **12**, 101831 (Jan 2025). https://doi.org/10.1016/j.ssaho.2025.101831
9. Jin, X., Zhou, W., Zhu, Q., Wang, W., Xu, G.: Research on the analysis and application of technological supply and demand structure based on LDA and BERTopic models. Cognitive Robotics **5**, 260–275 (Jan 2025). https://doi.org/10.1016/j.cogr.2025.07.001

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2

10. Kapantaidakis, I., Perakakis, E., Mastorakis, G., Kopanakis, I.: An Innovative Approach to Topic Clustering for Social Media and Web Data Using AI. Computers **14**(4), 142 (Apr 2025). https://doi.org/10.3390/computers14040142, publisher: Multidisciplinary Digital Publishing Institute
11. Lai, K.K., Hsu, Y.J., Chihwen, H.: Uncovering Mechanisms of Strategic Knowledge Evolution: Data-Driven Trajectories from BERTopic x PCC(1980-2025) (Oct 2025). https://doi.org/10.2139/ssrn.5616289
12. Li, C., Hu, X.: Medical Artificial Intelligence in Scholarly and Public Perspective: BERTopic-Based Analysis of Topic-Sentiment Collaborative Mining. Data Science and Informetrics (May 2025). https://doi.org/10.1016/j.dsim.2025.05.001
13. Ma, W., Ho, S.Y.: Sentiment-devoid lexicons: A novel method for domain-specific textual analysis in business and governance documents. Information & Management **62**(1), 104055 (Jan 2025). https://doi.org/10.1016/j.im.2024.104055
14. Mohamed, A.A.: Using BERTopic modelling to map the evolution of space syntax research. Land Use Policy **157**, 107639 (Oct 2025). https://doi.org/10.1016/j.landusepol.2025.107639
15. Obukhov, A., Krasnyanskiy, M., Nikolyukin, M.: Implementation of decision support subsystem in electronic document systems using machine learning techniques. In: 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). pp. 1–6 (2019). https://doi.org/10.1109/FarEastCon.2019.8934879
16. OECD: Smart City Data Governance: Challenges and the Way Forward. OECD Urban Studies (Oct 2023). https://doi.org/10.1787/e57ce301-en, publisher: OECD Publishing
17. Rachel J, J.L., A, B., M, K.: Topic Modeling Based Clustering of Disaster Tweets Using BERTopic. In: 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon). pp. 1–6 (Apr 2024). https://doi.org/10.1109/MITADTSoCiCon60330.2024.10575555
18. Son, H., Park, Y.E.: Agenda-setting effects for covid-19 vaccination: Insights from 10 million textual data from social media and news articles using BERTopic. International Journal of Information Management **83**, 102907 (Aug 2025). https://doi.org/10.1016/j.ijinfomgt.2025.102907
19. Walsh, J., Cave, J., Griffiths, F.: Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences. J Med Internet Res **26**, e54321 (Dec 2024). https://doi.org/10.2196/54321
20. Yang, C., Kim, Y.: Enhancing topic coherence and diversity in document embeddings using LLMs: A focus on BERTopic. Expert Systems with Applications **281**, 127517 (Jul 2025). https://doi.org/10.1016/j.eswa.2025.127517
21. Zadgaonkar, A., Agrawal, A.J.: An Approach for Analyzing Unstructured Text Data Using Topic Modeling Techniques for Efficient Information Extraction. New Generation Computing **42**(1), 109–134 (Mar 2024). https://doi.org/10.1007/s00354-023-00230-5
22. Švaňa, M.: Social Media, Topic Modeling and Sentiment Analysis in Municipal Decision Support (Aug 2023). https://doi.org/10.48550/arXiv.2308.04124, arXiv:2308.04124 [cs]
23. Švaňa, M., Zapletal, F., Hudec, M.: Citizen Engagement in Smart Cities Municipal Decision-Making. In: From Text to Understanding: Using Fuzzy Sets to Analyse Free-Form Text Data, pp. 1–17. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-032-00129-0_1

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_2