Classical Tricks for Better Topics: Can Traditional Text Mining Techniques Improve Generative Topic Modeling?

Isis Meira[1], Isabella Leite[1], Gabriel Assis[1],
Marcela Aniceto[2], Darian Rabbani[2], Lucas Pellicer[2], Aline Paes[1]

[1] Universidade Federal Fluminense, Instituto de Computação, Niterói, RJ, Brazil
{isismeira, isabellalps, assisgabriel}@id.uff.br, alinepaes@ic.uff.br
[2] Instituto de Ciência e Tecnologia Itaú, São Paulo, SP, Brazil
{marcela.aniceto-santos, darian.rabbani, lucas.pellicer}@itau-unibanco.com.br

Abstract. Topic modeling is widely used to discover and organize latent themes in large text collections. While recent embedding-based methods have improved topic discovery by leveraging dense semantic representations, they typically represent topics as lists of keywords, often providing limited insights into the underlying theme. More recently, Large Language Model (LLM)-based approaches, such as TopicGPT, have introduced generative topic modeling methods capable of producing human-readable topic labels and descriptions. However, the effectiveness of these methods depends heavily on the quality and representativeness of the documents selected to guide topic generation. In this paper, we propose Core-Periphery Sampling (CPS), a hybrid strategy that combines classical text representation and clustering methods with a TopicGPT-inspired pipeline. CPS selects both central documents and boundary instances that capture transitional and potentially informative cases, providing the generative model with an informed and diverse subset for topic induction. Experimental results show that CPS outperforms random sampling, particularly when the underlying document representation yields well-separable clusters. Although BERTopic remains stronger on clustering-oriented metrics, CPS offers an effective, computationally simple mechanism for guiding LLM-based topic generation via informed document selection, resulting in coherent and contextually grounded topic descriptions.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Machine learning algorithms**.

Keywords: clustering, document sampling, large language models, text mining, topic modeling

1. INTRODUCTION

Topic modeling aims to discover latent themes in large text collections and has become a central tool for organizing, exploring, and interpreting textual data at scale [Blei 2012]. This task is relevant to a wide range of applications, including recommendation systems, market and sentiment analysis, government auditing, scientific literature exploration, and social media monitoring [Abdelrazek et al. 2023]. Early approaches were largely dominated by statistical bag-of-words models, such as Latent Dirichlet Allocation (LDA) [Blei et al. 2003], which represent topics as distributions over words and documents as mixtures of topics.

More recently, contextualized embeddings and dense clustering methods have shifted the field toward semantic representations of documents [Meng et al. 2022]. Methods such as BERTopic [Grootendorst 2022] use Sentence-Transformer embeddings and clustering algorithms to group semantically related documents, often outperforming traditional lexical-based representations. However, topic interpre-

The authors acknowledge the support of CNPq (307088/2023-5); FAPERJ (SEI-260003/002930/2024, SEI-260003/000614/2023); CAPES (Finance Code 001), INCTs IAIA and TILD-IAR. The authors would also like to thank Instituto de Ciência e Tecnologia Itaú (ICTi). All the conclusions expressed by the authors do not reflect the opinions of Itaú Unibanco and Instituto de Ciência e Tecnologia Itaú. Also, this study must not result in any commercial process. Finally, all data used in this study comply with the Brazilian General Data Protection Law.

tation remains a significant challenge [Amorim et al. 2023]. Most embedding-based approaches still characterize topics through sets of representative keywords, such as '*court*, *judge*, and *appeal*'. Although these keywords provide clues about the underlying theme, they lack the contextual information necessary for a precise understanding of the topic. For instance, it may be unclear whether the theme concerns judicial decisions, legal institutions, criminal appeals, or another related aspect of the legal domain. As a result, users must often infer the semantic meaning of a topic themselves, which can limit the usefulness of topic modeling in real-world applications.

In contrast, the rise of Large Language Models (LLMs) has introduced a new paradigm for topic modeling based on prompt-driven generation. Frameworks such as TopicGPT [Pham et al. 2024] can synthesize documents into natural-language topic labels and descriptions, producing outputs that are often more informative and interpretable than keyword lists. Nevertheless, applying LLMs directly is not without challenges. Generative models are highly sensitive to prompt content, and poorly selected inputs may lead to vague, biased, or even hallucinated topic descriptions [Anh-Hoang et al. 2025]. Moreover, applying LLMs to large document collections is typically computationally expensive, especially when exploring configurations across prompts, models, and sampling strategies. In this context, we formulate the following research question:

> **RQ:** *How to cost-effectively select a representative subset of documents to improve the quality of topics generated by LLMs?*

To address this question, we propose Core-Periphery Sampling (CPS)[1]. CPS selects both representative documents located near cluster centroids and boundary documents situated at the interface between clusters. Our intuition is that central documents capture the dominant semantic characteristics of a topic, whereas peripheral documents reveal variations and ambiguities that might otherwise go unnoticed. Together, they provide the LLM with concise yet informative context for generating topics. To operationalize this idea, CPS combines classical text representation and clustering methods, such as TF-IDF [Manning et al. 2008], Latent Semantic Analysis (LSA) [Deerwester et al. 1990], and K-Means [MacQueen 1967], with a TopicGPT-inspired generative pipeline. In this way, the method aims to preserve the interpretability of generative topic modeling while using clustering structure to determine which documents should inform the generated taxonomy.

Our results on two datasets — Wiki [Merity et al. 2018] and Bills [Hoyle et al. 2021] — show that CPS can improve generative topic modeling by replacing random sampling with core-periphery document selection, especially when the underlying clustering structure is more separable. The gains are strongest for topic-name matching, while BERTopic remains competitive on clustering metrics but does not directly provide natural-language topic labels.

2. RELATED WORK

Topic modeling has traditionally been addressed through statistical and matrix-factorization methods. Latent Dirichlet Allocation (LDA) [Blei et al. 2003] models topics as probability distributions over words and documents as mixtures of topics, while Non-Negative Matrix Factorization (NMF) [Lee and Seung 2000] decomposes document-term matrices into non-negative components that can be interpreted as latent topics. Although these approaches are scalable and widely used, their outputs are usually represented as ranked word lists, which often require manual interpretation and may not clearly express the semantic scope of a topic [Amorim et al. 2023].

More recent embedding-based methods shift the focus from word co-occurrence patterns to dense semantic representations. Top2Vec [Angelov 2020] jointly embeds documents and words to discover topic vectors, while BERTopic [Grootendorst 2022] clusters Sentence-Transformer document embeddings and represents topics using class-based TF-IDF. These methods can capture semantic similarity

[1]All code, prompts and resources are available on GitHub: `https://github.com/Melllitau/Core-Periphery-Sampling`

beyond surface-level lexical overlap and often provide strong clustering performance. However, they still commonly describe topics through representative words rather than natural-language descriptions.

LLM-based topic models have recently emerged, with early work prompting LLMs to infer topics directly from text collections [Wang et al. 2023], studies investigating whether LLMs can serve as alternatives to traditional topic modeling pipelines [Mu et al. 2024], and approaches that use LLMs to complement neural topic models [Yang et al. 2025]. Among these methods, TopicGPT [Pham et al. 2024] is the closest to our work. It proposes a prompt-based framework in which an LLM first generates topic labels and descriptions from a sampled subset of documents, then refines redundant topics, and finally assigns the remaining documents to the resulting topic taxonomy. This design directly addresses the interpretability limitations of keyword-based topic models by producing natural-language topic names and descriptions.

Our method, CPS, builds on the TopicGPT pipeline while shifting the focus to document selection during topic generation. Whereas TopicGPT emphasizes the generation, refinement, and assignment stages, CPS uses classical text representation and clustering methods to select both central documents and boundary cases as input to the generative pipeline. This gives the model a more structured view of the corpus before inducing the topic taxonomy. In this sense, CPS complements LLM-based topic modeling, as it does not replace the generative pipeline but instead guides it through a core-periphery sampling strategy derived from the document space.

3. THE CORE-PERIPHERY SAMPLING METHOD

The Core-Periphery Sampling (CPS) method consists of three main stages: (§3.1) documents are grouped into clusters according to their similarity, computed from TF-IDF representations reduced with Latent Semantic Analysis (LSA); (§3.2) core and peripheral documents are identified within each cluster, capturing both highly representative instances and documents located near cluster boundaries; and (§3.3) the selected documents are used in a generative topic generation and assignment process. Figure 1 illustrates the overall pipeline. By guiding the generative step with a compact set of representative and boundary documents, the method encourages the generation of topics that reflect both the central themes of each cluster and the semantic transitions between them.

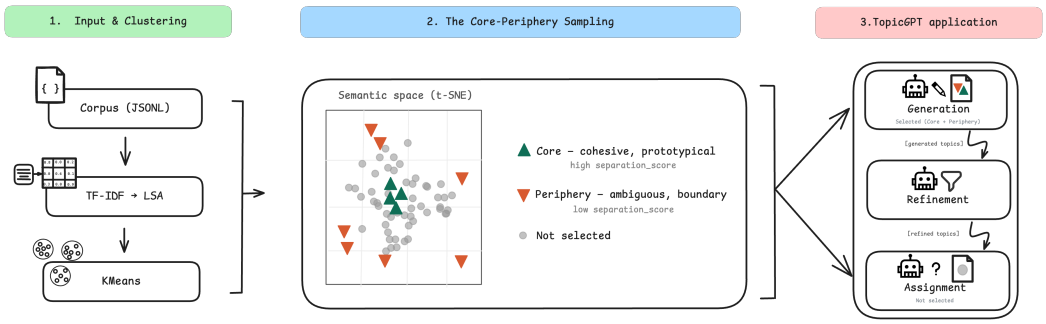


Fig. 1: The Core-Periphery Sampling method overview.

3.1 Clustering Process

In the clustering stage, CPS first represents documents using TF-IDF features [Manning et al. 2008] and projects them into a lower-dimensional semantic space with Latent Semantic Analysis (LSA) [Deerwester et al. 1990]. The resulting vectors are then clustered with K-Means [MacQueen 1967], grouping documents according to their lexical-semantic similarity. This step provides the structure for later identification of core and peripheral documents. We deliberately adopt TF-IDF and

LSA instead of more recent embeddings to maintain a lightweight, interpretable, and model-agnostic pipeline. Since CPS is intended as a sampling strategy that can be applied independently of the downstream LLM used for topic generation, relying on classical representations avoids introducing biases from a specific embedding model while keeping computational costs low and clustering behavior easier to analyze.

Concretely, documents are represented with unigram and bigram TF-IDF features, which are projected into a lower-dimensional latent semantic space through TruncatedSVD to obtain their LSA representation [Deerwester et al. 1990]. This transformation produces a compact representation of the original sparse document-term matrix, reducing lexical noise while preserving the main patterns of variation across documents. The resulting vectors are then L2-normalized[Manning et al. 2008], making document comparisons less sensitive to vector magnitude. K-Means is evaluated with different values of (K), and the final number of clusters is selected according to the silhouette score [Rousseeuw 1987]. The resulting TF-IDF/LSA representation defines the semantic space used by CPS to distinguish documents that are central to a cluster and those located near cluster boundaries.

3.2 The Core-Periphery Selection

After clustering, CPS assigns each document a separation score defined as the difference between its average distance to documents in other clusters and its average distance to documents in the same cluster. Formally, for a document d_i, the score is computed as $s_i = inter_i - intra_i$, where $intra_i$ denotes the mean intra-cluster distance and $inter_i$ denotes the mean inter-cluster distance in the LSA-reduced space. Higher scores indicate cohesive documents, which are closer to their own cluster and farther from others, while lower scores indicate ambiguous documents, which are positioned closer to cluster boundaries. This distinction allows CPS to sample both representative documents and boundary cases, preserving the central structure of each cluster as well as semantic transitions between clusters.

For document selection, CPS combines cohesive and ambiguous documents within each cluster. In this work, we adopt a balanced 50%/50% split between these two groups as the default configuration. For less compact clusters, identified by an average intra-cluster distance above the global average, this ratio is adjusted to 30% cohesive and 70% ambiguous documents. These proportions are configurable parameters of the method, allowing the sampling strategy to be adapted to different collections. The adopted configuration increases the presence of boundary cases in more dispersed clusters while retaining central examples from each cluster. As a result, the selected subset provides a structured view of the collection, guiding topic generation with documents that capture both dominant themes and peripheral variations.

3.3 TopicGPT application

We use TopicGPT [Pham et al. 2024] as the generative component of CPS because it separates topic discovery from topic assignment, allowing the model to first induce an interpretable topic taxonomy and then apply it consistently to the full collection. This is particularly suitable for CPS, since the generation phase can be guided by the selected core and peripheral documents, while the remaining documents are later assigned to the refined topic taxonomy.

The TopicGPT pipeline consists of three phases: generation, refinement, and assignment. During generation, the documents selected by CPS are presented individually to an LLM, which identifies high-level topics and incrementally builds a topic taxonomy by either assigning documents to existing topics or creating new ones. In the refinement phase, the generated topics are compared in pairs, allowing the LLM to merge semantically equivalent or highly similar topics and reduce redundancy in the taxonomy. Finally, during assignment, all remaining documents are classified into the refined topic taxonomy, with the LLM selecting the most appropriate existing topic for each document without introducing new categories.

4. EXPERIMENTAL RESULTS

This section describes the experimental setup used to implement and evaluate CPS, including the datasets, model configurations, baselines, hyperparameters, and evaluation metrics. Next, it presents the results obtained with CPS and the selected baselines.

4.1 Experimental Setup

Datasets. We evaluate CPS on two publicly available topic-labeled datasets. First, Wiki [Merity et al. 2018] consists of Wikipedia passages organized into 15 top-level topics and 45 subtopics, totaling 14,290 documents. Second, Bills [Hoyle et al. 2021] contains U.S. Congressional bill summaries spanning 21 top-level topics and 112 subtopics, totaling 32,661 documents. For both datasets, we remove instances without topic labels and discard documents associated with conflicting labels. After this preprocessing step, Wiki contains 14,235 documents and Bills contains 26,028 documents.

CPS Setup. Documents are represented using TF-IDF with the top 1,000 unigram and bigram features. Stop words are removed using a multilingual list composed of Portuguese, English, and Spanish terms. The resulting TF-IDF matrix is projected into a 50-dimensional LSA space using TruncatedSVD and then L2-normalized. K-Means is evaluated for $K \in [5, 40]$, with $n\_init = 10$ and $random\_state = 42$. The final number of clusters is selected according to the silhouette score, computed on a sample of 2,000 documents.

We use models from the Qwen-3 [Alibaba Cloud 2025] family to instantiate TopicGPT. The 14B variant is used for generation, while the 4B variant is used for assignment. This follows the original TopicGPT recommendation of using a larger model for the more open-ended generation stage and a smaller model for assignment to improve efficiency. The Qwen-3 family was selected due to its strong performance among open-weight models[2] and its compatibility with the computational resources available for our experiments. All remaining hyperparameters, including the number of documents used in the generation stage, as well as the prompts used in each phase, follow the original TopicGPT setup.

Baselines. CPS is compared with two reference methods. The TopicGPT baseline follows the original TopicGPT document selection strategy, randomly sampling 1,020 documents for the topic generation stage. The same LLMs, prompts, and remaining pipeline components are used as in CPS, with document selection being the only difference between the methods. This setup isolates the impact of the proposed core-periphery sampling strategy from that of the underlying language models and generation process. We also compare CPS against BERTopic [Grootendorst 2022], a neural topic modeling framework that induces topics by clustering Sentence-Transformer document embeddings, serving as a strong non-generative baseline. BERTopic is run with its default hyperparameters.

Evaluation Setup. We evaluate topic quality by measuring the alignment between predicted topic assignments and ground-truth labels. Following TopicGPT, external clustering metrics are used to compare predicted clusters against gold classes. *Purity* [Zhao and Karypis 2001] measures how consistently each ground-truth category is associated with its best-matching predicted cluster. *Adjusted Rand Index* (*ARI*) [Hubert and Arabie 1985; Vinh et al. 2010] measures pairwise agreement between two clusterings while correcting for chance. *Normalized Mutual Information* (*NMI*) [Strehl and Ghosh 2003] measures the amount of shared information between predicted and gold clusters, normalized to reduce sensitivity to the number of clusters. In addition, *Topic Alignment* [Chuang et al. 2013] combines *Purity* and *NMI*, while *Topic Coherence* (C_V) [Röder et al. 2015] measures the semantic consistency of the most representative words in each topic based on co-occurrence patterns.

[2] `https://artificialanalysis.ai/models/qwen3-14b-instruct-reasoning`

For the TopicGPT variants, we also evaluate how well the generated topic names match the ground-truth labels. *Exact Match* measures the percentage of identical labels, while *Cosine Similarity* compares their sentence-level embeddings. *BERTScore F1* [Zhang et al. 2020] evaluates semantic similarity with contextual embeddings, *Jaccard Similarity* measures word overlap, and *Levenshtein Similarity* captures character-level similarity through edit distance [Manning et al. 2008]. These metrics are not computed for BERTopic, since it does not generate natural-language topic labels directly.

4.2 Results

Table I reports the results for topic modeling evaluation, separating metrics related to topic-labeling matching (*Name*) from those related to document-topic assignment (*Topic*). TopicGPT and BERTopic are reported with standard deviations over three runs, while CPS is reported without deviation because the fixed hyperparameter configuration and selected models produced the same outputs across runs. To assess the stability of the differences between our method and the original TopicGPT, 95% bootstrap confidence intervals (CIs) are reported for TopicGPT, CPS, and their difference Δ, using $B = 10{,}000$ paired resamples over documents, except for Topic Coherence, where topics are resampled.

Table I: Comparison between BERTopic, the original TopicGPT pipeline with random sampling, and our CPS method. The Δ denotes CPS minus TopicGPT. Best results are shown in **bold**.

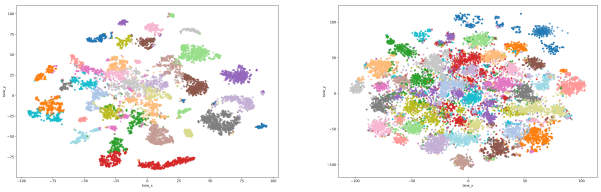
Type	Metric	BERTopic	TopicGPT		CPS (ours)		CPS − TopicGPT	
			Score	95% CI	Score	95% CI	Δ	95% CI
Wiki								
Name	Exact Match (%)	–	1.6673 ± 2.6078	[1.5014, 1.7821]	**12.3647**	[11.7238, 12.9731]	+**10.6974**	[10.0713, 11.3550]
	Cosine Similarity	–	0.2925 ± 0.0881	[0.2899, 0.2937]	**0.4063**	[0.4029, 0.4100]	+**0.1138**	[0.1109, 0.1181]
	BERTScore F1	–	0.7115 ± 0.0932	[0.7050, 0.7135]	**0.8696**	[0.8688, 0.8704]	+**0.1581**	[0.1561, 0.1647]
	Jaccard Similarity	–	0.0638 ± 0.0616	[0.0624, 0.0650]	**0.1249**	[0.1208, 0.1292]	+**0.0611**	[0.0568, 0.0657]
	Levenshtein Similarity	–	0.1671 ± 0.0447	[0.1652, 0.1680]	**0.2645**	[0.2607, 0.2676]	+**0.0974**	[0.0939, 0.1013]
Topic	Harmonic Mean Purity	**0.6198** ± 0.0061	0.3567 ± 0.0328	[0.3542, 0.3631]	0.4260	[0.4188, 0.4358]	+**0.0693**	[0.0608, 0.0766]
	Adjusted Rand Index	**0.4465** ± 0.0082	0.1464 ± 0.0806	[0.1431, 0.1517]	0.2363	[0.2251, 0.2436]	+**0.0899**	[0.0783, 0.0955]
	Normalized Mutual Info.	**0.5911** ± 0.0058	0.2764 ± 0.0914	[0.2734, 0.2832]	0.3629	[0.3557, 0.3715]	+**0.0865**	[0.0776, 0.0933]
	Topic Coherence	**0.6536** ± 0.0182	0.5293 ± 0.0211	[0.4782, 0.5649]	0.5576	[0.4901, 0.6099]	+**0.0283**	[−0.0461, 0.1034]
	Topic Alignment	**0.6055** ± 0.0044	0.3166 ± 0.0621	[0.3141, 0.3228]	0.3944	[0.3880, 0.4030]	+**0.0778**	[0.0699, 0.0842]
Bills								
Name	Exact Match (%)	–	7.7002 ± 8.0263	[7.5026, 7.9356]	**9.5809**	[9.1159, 9.8917]	+**1.8807**	[1.4810, 2.0929]
	Cosine Similarity	–	**0.3986** ± 0.0188	[0.3973, 0.4013]	0.3746	[0.3717, 0.3776]	-0.0240	[−0.0270, −0.0222]
	BERTScore F1	–	**0.8764** ± 0.0136	[0.8758, 0.8778]	0.8645	[0.8628, 0.8665]	-0.0119	[−0.0137, −0.0107]
	Jaccard Similarity	–	**0.1102** ± 0.0482	[0.1084, 0.1128]	0.1021	[0.0987, 0.1049]	-0.0081	[−0.0112, −0.0063]
	Levenshtein Similarity	–	0.2338 ± 0.0463	[0.2325, 0.2359]	**0.2505**	[0.2474, 0.2528]	+**0.0168**	[0.0135, 0.0184]
Topic	Harmonic Mean Purity	**0.4461** ± 0.0136	0.4010 ± 0.0301	[0.3975, 0.4063]	0.3760	[0.3704, 0.3836]	-0.0250	[−0.0313, −0.0186]
	Adjusted Rand Index	0.1795 ± 0.0190	**0.2048** ± 0.0417	[0.2011, 0.2109]	0.1866	[0.1813, 0.1921]	-0.0182	[−0.0245, −0.0141]
	Normalized Mutual Info.	**0.3886** ± 0.0112	0.3641 ± 0.0217	[0.3644, 0.3733]	0.3625	[0.3617, 0.3729]	-0.0017	[−0.0062, 0.0031]
	Topic Coherence	0.4817 ± 0.0058	0.4800 ± 0.0115	[0.4597, 0.4948]	**0.4939**	[0.4656, 0.5221]	+**0.0139**	[−0.0170, 0.0500]
	Topic Alignment	**0.4173** ± 0.0120	0.3826 ± 0.0246	[0.3812, 0.3896]	0.3693	[0.3667, 0.3777]	-0.0133	[−0.0181, −0.0083]

The differences between CPS and the original TopicGPT pipeline are significant according to the CI analysis, except for Topic Coherence and Bills NMI. On Wiki, CPS consistently outperforms the original TopicGPT pipeline across all metrics. Since both methods share the same generative pipeline, prompts, and underlying models, these gains can be attributed to the proposed core-periphery selection, which appears to provide a more informative subset than random sampling. On Bills, the results are more stable across methods, with CPS and TopicGPT obtaining comparable performance across most metrics. Nevertheless, across both datasets, the largest gains of CPS over TopicGPT are observed in topic-name Exact Match, suggesting that core-periphery selection is particularly useful for guiding the generation of topic labels that more directly match the human ground-truth taxonomy.

BERTopic achieves the best results across most clustering metrics, especially on Wiki. This contrasts with the original TopicGPT findings [Pham et al. 2024], where the generative approach outperformed BERTopic; however, their setup relied on more powerful closed-source models, namely GPT-4, for topic generation. These results indicate that neural non-generative topic models remain competitive

for document clustering. However, BERTopic does not directly generate natural-language topic names, which limits its usefulness in applications that require interpretable topic labels.

Figure 2 shows the t-SNE [van der Maaten and Hinton 2008] visualization of the LSA-reduced document space used before core-periphery selection. Wiki (Figure 2a) presents clearer cluster separation than Bills (Figure 2b), suggesting that CPS benefits more when the underlying vector-space representation yields well-separated groups. This helps explain why CPS clearly outperforms random TopicGPT on Wiki, while both methods remain close on Bills.



(a) Wiki (b) Bills

Fig. 2: t-SNE visualization of the document representations before CPS for the Wiki and Bills datasets.

Overall, regarding our main RQ, the results indicate that CPS, a document selection strategy built upon classical text representation techniques, constitutes an effective alternative for supporting LLM-based topic generation. Compared to TopicGPT's random sampling strategy, CPF offers consistent gains under favorable clustering conditions while preserving the advantage of generative methods in producing interpretable topic labels. Although BERTopic achieves stronger results on most clustering metrics, it does not directly generate natural-language topic names, which remains an important limitation when the goal is to obtain interpretable topic labels aligned with a target taxonomy. Thus, CPS occupies a complementary position: in our setup, it does not always outperform neural clustering methods in assignment, but it improves the generative pipeline where labeling is central to the task.

5. CONCLUSIONS AND FUTURE WORK

In this work, we proposed CPS, a hybrid topic modeling strategy that leverages the structure of the document space to guide generative topic modeling. CPS combines classical text representations and clustering methods with a TopicGPT-inspired pipeline that selects both central documents and boundary cases as input for topic generation. In doing so, CPS investigates whether classical text mining techniques can help select more informative documents for generative topic modeling.

Our experiments on two datasets, Wiki and Bills, show that CPS can improve over random sampling in the TopicGPT pipeline, especially when the underlying clustering structure is more separable. At the same time, the results also show that BERTopic remains competitive, particularly on clustering-based metrics. These findings indicate that CPS occupies a complementary position: it improves the document selection stage of generative topic modeling while preserving the advantages of LLM-generated topic labels, which provide explicit topic names and descriptions.

This work also has limitations that point to future directions. We instantiate the generative pipeline with a single LLM family, Qwen-3, and do not explore a broad range of model sizes, prompts, or decoding configurations. CPS, built on K-means clustering and silhouette-based selection, may also behave differently for non-convex or variable-density cluster structures, although preliminary experiments with methods such as HDBSCAN [Campello et al. 2015] suggest that it remains effective in such settings. Future work should therefore evaluate CPS across different LLM families and clustering methods. In addition, our experiments use fixed CPS proportions for selecting central and boundary documents; a broader analysis of these proportions and related hyperparameters could clarify when core-periphery sampling is most effective. Finally, our evaluation focuses mostly on English datasets, motivating further analysis in other languages and multilingual corpora.

REFERENCES

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. Topic modeling algorithms and applications: A survey. *Information Systems* vol. 112, pp. 102131, 2023.

Alibaba Cloud. Qwen3 technical report, 2025.

Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D., and Paes, A. ALTES: uma ferramenta de rotulação automática de tópicos por meio de fontes externas. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*. SBC, Porto Alegre, RS, Brasil, pp. 120–125, 2023.

Angelov, D. Top2vec: Distributed representations of topics, 2020.

Anh-Hoang, D., Tran, V., and Nguyen, L.-M. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence* vol. 8, pp. 1622292, 2025.

Blei, D. M. Probabilistic topic models. *Commun. ACM* 55 (4): 77–84, Apr., 2012.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* vol. 3, Mar., 2003.

Campello, R., Moulavi, D., Zimek, A., and Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data* 10 (1): 1–51, 2015.

Chuang, J., Gupta, S., Manning, C., and Heer, J. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. 28. PMLR, Atlanta, Georgia, USA, pp. 612–620, 2013.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 (6): 391–407, 1990.

Grootendorst, M. BERTopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., and Resnik, P. Is automated topic model evaluation broken? the incoherence of coherence. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Curran Associates Inc., Red Hook, NY, USA, 2021.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification* 2 (1): 193–218, Dec, 1985.

Lee, D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press, 2000.

MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. UC Press, Berkeley, CA, 1967.

Manning, C., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge UP, 2008.

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., and Han, J. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM web conference 2022*. pp. 3143–3152, 2022.

Merity, S., Keskar, N. S., and Socher, R. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*, 2018.

Mu, Y., Dong, C., Bontcheva, K., and Song, X. Large language models offer an alternative to the traditional approach of topic modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, pp. 10160–10171, 2024.

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Mexico City, Mexico, pp. 2956–2984, 2024.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Association for Computing Machinery, New York, NY, USA, pp. 399–408, 2015.

Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* vol. 20, pp. 53–65, 1987.

Strehl, A. and Ghosh, J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3 (null): 583–617, Mar., 2003.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* 11 (95): 2837–2854, 2010.

Wang, H., Prakash, N., Hoang, N. K., Hee, M. S., Naseem, U., and Lee, R. K.-W. Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)*. pp. 1236–1241, 2023.

Yang, X., Zhao, H., Xu, W., Qi, Y., Lu, J., Phung, D., and Du, L. Neural topic modeling with large language models in the loop. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, pp. 1377–1401, 2025.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Zhao, Y. and Karypis, G. Criterion functions for document clustering, 2001.