Anchored Topic Modeling for Interpretable Clinical Trial Landscapes

Piotr Eliasz
Department of Computer Science
Aarhus University
eliasz.piotr@icloud.com

Mate Kornidesz
Department of Computer Science
Aarhus University
kornimate@gmail.com

Peeter Tarvas
Department of Computer Science
Aarhus University
peetertarvas@gmail.com

Abstract

The growth of clinical research has made it difficult to track how medical topics evolve across thousands of trial protocols. Traditional topic modeling techniques often lack interpretability in complex biomedical domains. This project develops an **Anchored Topic Modeling** framework combining transformer-based text embeddings with structured metadata from the **Clinical Trials** dataset. By utilizing domain-specific anchors derived from MeSH terms, we guide topic formation towards medically meaningful categories while preserving the model's ability to discover novel latent themes. Our results demonstrate that this semi-supervised approach produces a structured, interpretable representation of the clinical research landscape, bridging the gap between unsupervised text analysis and expert domain knowledge.

1 Introduction

The rapid increase of clinical trials has resulted in a huge amount of unstructured text data. With hundreds of thousands of trials on platforms like ClinicalTrials.gov, it is difficult for researchers to manually monitor new trends. The core value of this textual data lies in free-text fields like `brief_summary` and `detailed_description`. They contain rich semantic information that is difficult to analyze at scale using keyword-based search.

Classic approaches like Latent Dirichlet Allocation (LDA) rely on bag-of-words representations, which are unable to capture semantic nuances (e.g., the distinction between “cold” as a temperature and “cold” as a virus). In contrast, modern unsupervised neural approaches (e.g., BERT) excel at semantic similarities. However, they often lack **interpretability** because they generate clusters based on procedural similarities rather than therapeutic domains.

To address this issue, we propose the **Anchored Topic Modeling** framework. We combine the semantic potential of **Bio_ClinicalBERT** with the structural rigor of expert-developed metadata. Using a semi-supervised **Spherical K-Means** algorithm with “anchors” (derived from medical taxonomy), we can tailor the model to specific domains while maintaining flexibility in classifying unseen data.

Our contributions are:

- A hybrid representation pipeline integrating unstructured clinical text with structured metadata to optimize medical interpretability.

- An anchor-guided clustering mechanism that overcomes the "cold start" problem of unsupervised learning.
- An interface capable of mapping raw clinical descriptions to interpretable medical topics.

2 Related Work

2.1 Topic Modeling Evolution

Latent Dirichlet Allocation (LDA) (1) has long been the standard solution for extracting topics from text. However it is limited by the “bag-of-words” assumption. More recent approaches, such as BERTopic (2), use transformer embeddings for geometric document clustering. Although this method is effective, unsupervised methods often tend generate topics based on statistics rather than expert taxonomy.

2.2 Domain-Specific Embeddings

General-purpose language models often underperform on specialized text data due to domain shift. To mitigate this problem, models such as ClinicalBERT (3) can be used. They have been specifically tailored to handle the biomedical domain. Produced embeddings ensure that the latent space reflects medical ontology rather than general linguistic proximity.

2.3 Anchored and Seeded Clustering

To better adapt to specialized knowledge, it is possible to use semi-supervised approaches that take into account previous limitations. Our approach is based on the Seeded K-Means method (4), in which the optimization process is guided by predefined centroids. This ensures that the resulting clusters correspond to recognized medical categories and the data structure allows for the definition of detailed subtopics.

3 Methodology

In our work we propose a hybrid framework designed to structure the semantic landscape of clinical trials. The pipeline consists of four stages: Data Construction, Neural Representation, Anchor-Guided Clustering and Inference.

3.1 Data Input and Preprocessing

The primary data source is the Clinical Trials dataset[1].

For each trial t, we construct a single dense input sequence S_t by concatenating unstructured fields (`brief_summary`, `detailed_description`) with structured descriptors (`mesh_terms`, `keywords`). To mitigate noise, we apply a preprocessing pipeline:

- **Text Cleaning:** Removal of HTML tags and special characters via RegEx.
- **Linguistic Normalization:** Using `spaCy`, we perform stop-word removal and lemmatization to reduce vocabulary sparsity.
- **Tokenization:** Sequences are truncated to $L = 256$ tokens to fit the Bio_ClinicalBERT context window.

3.2 Neural Representation Architecture

We employ **Bio_ClinicalBERT** (3) as our backbone encoder. This model utilizes the standard *BERT-Base* architecture ($L = 12$ layers, $H = 768$ hidden size, $A = 12$ attention heads) to process the tokenized clinical sequences. By utilizing its domain-specific pre-training, the encoder effectively distinguishes ambiguous medical terms based on syntactic context.

[1] The data used in this project is sourced from the clinical-trials dataset, available on the Hugging Face platform: `https://huggingface.co/datasets/louisbrulenaudet/clinical-trials`

3.2.1 Embedding Strategy

The transformer outputs a sequence of contextualized vectors. We derive a fixed-size document representation via **Mean Pooling** over the valid token vectors:

$$v_{raw} = \frac{\sum_{i=1}^{L} H_i \cdot M_i}{\sum_{i=1}^{L} M_i + \epsilon} \tag{1}$$

where H_i represents the hidden state of the i-th token and ϵ ensures numerical stability.

In high-dimensional spaces (768 dimensions), standard Euclidean distance becomes less meaningful due to the dimensionality problem. To mitigate this, we apply **L2-normalization**, projecting every document embedding onto the unit hypersphere (S^{d-1}).

$$v = \frac{v_{raw}}{||v_{raw}||_2} \tag{2}$$

This ensures that clustering is driven purely by the **semantic angle** rather than vector magnitude, making the subsequent Spherical K-Means robust to varying text lengths.

3.3 Anchor-Guided Spherical Clustering

Building upon the spherical representation established in the previous stage, we implement **Seeded Spherical K-Means**. This approach leverages the directional properties of the normalized embeddings to perform clustering that is explicitly guided by expert knowledge.

3.3.1 Mathematical Formulation

Since all document vectors are projected onto the unit hypersphere (S^{d-1}), the standard Euclidean distance metric inherently adapts to measuring angular proximity. We rely on the property that minimizing Euclidean distance on unit-length vectors is mathematically equivalent to maximizing cosine similarity (5).

For any normalized data point x and centroid c (where $\|x\| = \|c\| = 1$), this relationship is derived as follows:

$$\begin{aligned} \|x - c\|^2 &= (x - c) \cdot (x - c) \\ &= x \cdot x + c \cdot c - 2(x \cdot c) \\ &= 1 + 1 - 2(x \cdot c) \\ &= 2\big(1 - \text{cosine\_similarity}(x, c)\big) \end{aligned} \tag{3}$$

This equivalence ensures that the K-Means optimization objective naturally partitions the data based on semantic alignment.

3.3.2 Algorithm Configuration and Initialization

We configured the algorithm with $K = A + E$ clusters. Specifically:

- $A = 17$: Fixed anchor centroids derived from expert topics (MeSH terms).
- $E = 20$: Randomly initialized centroids to capture latent sub-topics.

This results in a total of $K = 37$ clusters.

3.4 Inference Interface

To utilize the framework, we developed a lightweight inference engine designed for real-time semantic tagging. This module acts as a bridge between raw text and medical insights. When a new clinical trial description is input, the interface first processes the text through our cleaning pipeline to standardize the language. Next, it converts the text into a mathematical representation using our pre-trained model. Finally, it compares this representation against the established cluster centers to identify the best match. Interface returns the most relevant medical topic along with its interpretative keywords. This design allows for immediate categorization of new research proposals without requiring model retraining.

4 Experiments and Evaluations

To validate the efficacy of the framework, we assessed both geometric compactness and semantic alignment.

4.1 Experimental Setup

The experimental pipeline was implemented using `PyTorch` for feature extraction and `scikit-learn` for clustering operations.

4.1.1 Data Partitioning and Encoding

We utilized the **full processed dataset** of clinical protocols. The text was encoded using **Bio_ClinicalBERT** in inference mode (weights frozen). We utilized a batch size of 128 to optimize GPU memory usage, resulting in a dense feature matrix $X \in \mathbb{R}^{N \times 768}$. The data was stratified into an 80% training set ($N_{train} \approx 383k$) used for clustering optimization and a 20% hold-out test set ($N_{test} \approx 96k$) for evaluation.

4.1.2 Clustering Configuration

The Seeded Spherical K-Means algorithm was configured with $K = 37$ clusters. This number was explicitly composed of $A = 17$ fixed anchor centroids (derived from high-density MeSH terms) and $E = 20$ randomly initialized centroids to capture latent sub-topics. To ensure convergence, we set the maximum iterations to 300.

4.2 Evaluation Metrics

We employed a dual-metric strategy:

4.2.1 Geometric Metrics

- **Silhouette Score:** Measures cluster cohesion and separation (-1 to +1).
- **Calinski-Harabasz (CH) Index:** Ratio of between-cluster to within-cluster dispersion. Higher values indicate denser, better-separated clusters.
- **Davies-Bouldin (DB) Index:** Average similarity between each cluster and its most similar one (lower is better).

4.2.2 Interpretability Metrics

- **Anchor Coverage:** Proportion of documents with similarity above a threshold (0.5) to at least one anchor.
- **Cluster Anchor Purity:** Probability weight of the dominant anchor within a cluster. High purity implies exclusive association with a single medical domain.
- **Global NMI:** Alignment between cluster assignments and dominant anchor labels.

5 Results and Discussion

In this section, we evaluate the stability and interpretability of the proposed framework. We first analyze the impact of hyperparameters (τ, K) using an ablation study, and then report the performance of the final configured model. Due to the limitation of the testing here is done with a subsample of data 100 000 which took around 22-25 minutes per testing round. This is important because in the end we needed to decide on what parameters we will use in the pipeline with the full dataset, as it took long time to train and optimizing with the full dataset would thus not be optimal.

5.1 Impact of Anchor Sharpness (τ)

Table 1 quantifies the relationship between the temperature parameter τ and the model's ability to map documents to expert-defined domains. We observe a critical phase transition between $\tau = 0.3$ and

$\tau = 0.1$. Reducing the temperature to $\tau = 0.1$ minimizes the entropy of the assignment distribution, effectively acting as a **semantic contrast filter**. This "sharpening" yielded the highest Purity and Anchor Coverage as was expected, confirming that a low-temperature regime is required to bridge the gap between continuous vector spaces and categorical medical terminology.

Table 1: Effect of temperature τ on anchor-alignment metrics. Lower τ values force "harder" decisions, significantly improving semantic alignment without affecting geometric structure.

Parameter	Sil.	CH Score	DB Index	Align (%)	Purity	Lift	NMI
$\tau = 0.9$	0.0221	433.55	3.2085	0.0	0.078	0.0184	0.2186
$\tau = 0.7$	0.0221	433.55	3.2085	0.0	0.084	0.0245	0.2186
$\tau = 0.5$	0.0221	433.55	3.2085	0.0	0.097	0.0367	0.2186
$\tau = 0.3$	0.0221	433.55	3.2085	0.7	0.132	0.0691	0.2186
$\tau = 0.1$	**0.0221**	**433.55**	**3.2085**	**26.2**	**0.296**	**0.2128**	**0.2186**

5.2 Impact of Cluster Granularity (K)

able 2 investigates the trade-off between geometric compactness and topic granularity during our preliminary tuning. We observed that **structural saturation** between $E = 100$ and $E = 120$, where metrics remained identical, suggesting that the fixed anchors dominate the optimization landscape at higher $K = (A + E)$. Lowering the E, the metrics started slowly changing for the better.

Table 2: Ablation study on cluster size E (at $\tau = 0.1$). Increasing E led to structural saturation, prompting the selection of a smaller, more compact K=37 for the final model.

Configuration	Sil.	CH Score	DB Index	Align (%)	Purity	Lift	NMI
$E = 20$	0.0267	1166.76	2.9634	26.2	0.315	0.2251	0.2020
$E = 40$	0.0249	842.93	3.0228	26.2	0.319	0.2367	0.2151
$E = 60$	0.0253	676.11	3.0793	26.2	0.304	0.2224	0.2129
$E = 80$	0.0238	564.60	3.0592	26.2	0.303	0.2205	0.2175
$E = 100$	0.0221	433.55	3.2085	26.2	0.296	0.2128	0.2186
$E = 120$	0.0221	433.55	3.2085	26.2	0.296	0.2128	0.2186

In this section, we evaluate the stability and interpretability of the proposed framework. We trained the model on the full dataset of 479,038 rows, part of which was used for testing approximately 96,000 rows - 20 percent train test split. The parameters that were chosen in the previous test ($K = 37, \tau = 0.1$).

5.3 Geometric Compactness

The model demonstrated strong structural coherence in the high-dimensional space.

- **Silhouette Score:** 0.0272. This positive value confirms that on average, documents are closer to their assigned centroids than to neighboring clusters.
- **Calinski-Harabasz (CH) Index:** 1473.82. This significantly high score indicates dense and well-separated clusters, validating the effectiveness of using the full dataset compared to smaller subsets.
- **Davies-Bouldin (DB) Index:** 3.3148.

5.4 Semantic Alignment and Interpretability

Beyond geometry, the model aligned unstructured text with medical taxonomy.

- **Anchor Coverage:** 25.8%. Approximately one-quarter of all trials strongly aligned (similarity > 0.5) with the predefined medical anchors, suggesting the model effectively captured some of the core medical domains.
- **Cluster Anchor Purity:** 0.285. This indicates that the clusters maintained a consistent semantic theme relative to the expert anchors.

- **Global NMI:** 0.1995. This score reflects a balanced alignment between the unsupervised cluster assignments and the dominant anchor labels.

These results confirm that the **Anchored Topic Modeling** framework successfully structures the clinical trial landscape, creating clusters that are both geometrically dense and medically interpretable.

6 Conclusion and Future Work

In this work, we presented an **Anchored Topic Modeling** framework designed to provide structure and interpretability to clinical trial protocols. By integrating the semantic richness of **Bio_ClinicalBERT** with the structural guidance of expert-defined anchors, we bridged the gap between unsupervised discovery and medical taxonomy.

Our experiment showed that initializing spherical clustering with domain-specific centroids creates a strong inductive bias which aligns latent topics with medical fields. Configuration ($K = 37$) achieved best geometric density (Calinski-Harabasz Score > 1470) while preserving semantic purity. The proposed inference interface further utilizes this model, allowing for semantic tagging of new research proposals.

6.1 Future Work

There are several areas for the future framework improvement. First, computational efficiency could be increased by optimizing the text processing pipeline to handle larger datasets more quickly. Second, the coverage of medical topics could be expanded by adding more synonyms from external medical dictionaries to the anchor list. Finally, a hierarchical approach could be implemented to automatically discover detailed sub-topics within the main medical categories.

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine learning research*, 3(Jan), 993-1022.

[2] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

[3] Alsentzer, E., Murphy, J. R., Boag, W., et al. (2019). Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72-78.

[4] Basu, S., Banerjee, A., & Mooney, R. (2002). Semi-supervised clustering by seeding. *In Proceedings of the 19th International Conference on Machine Learning (ICML-02)*, 27-34.

[5] Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1), 143-175.

[6] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

[7] Huang, J., Gong, S., & Zhu, X. (2020). Deep semantic clustering by partitioning high order neighbors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8837-8846.