MDS4AGT
Abstract Classification
SS25

Scenario 3:
Medical Abstract Classification using LLMs

Sara Farshi, Narges Shafieyoun, Yasser Ibourk

21.07.2025

Content

1. Introduction

The rapid expansion of biomedical literature presents a major challenge for healthcare professionals and researchers seeking timely and relevant information. Automated classification of medical abstracts provides a scalable solution for organizing this growing body of knowledge, with applications ranging from clinical decision support to systematic literature reviews and pharmacovigilance (Sakai, H., & Lam, S. S., 2025).

This project tackles the task of categorizing medical abstracts into five disease-related categories: *Neoplasms*, *Digestive System Diseases*, *Nervous System Diseases*, *Cardiovascular Diseases*, and *General Pathological Conditions*. The dataset used, "TimSchopf/medical_abstracts", is sourced from the Hugging Face repository and includes over 14,000 labeled abstracts split into training and test sets.

The main goal of this scenario is to design and evaluate a robust NLP pipeline using large language models (LLMs) for text classification in the medical domain. Specifically, we implement and compare three distinct methodological paradigms:

- **Classification using pre-trained representation models**, with and without task-specific fine-tuning,
- **Embedding-based classification**, including both supervised and zero-shot setups, and
- **Clustering and topic modeling** using dimensionality reduction and unsupervised techniques.

Each approach is analyzed in terms of classification accuracy, interpretability, and computational efficiency. The implementation leverages modern machine learning libraries such as Hugging Face Transformers (Wolf et al., 2020) and SentenceTransformers to build reproducible and scalable solutions tailored to the complexity of medical language.

2. Hugging Face Dataset

The medical abstract dataset used in this project is sourced from the Hugging Face Hub under the name "TimSchopf/medical_abstracts", originally presented in the publication *"Evaluating Unsupervised Text Classification: Zero-Shot and Similarity-Based Approaches"* (NLPIR 2022). The dataset consists of 14,438 scientific abstracts from the biomedical domain, manually categorized into five major disease-related classes:

1. Neoplasms
2. Digestive System Diseases
3. Nervous System Diseases
4. Cardiovascular Diseases
5. General Pathological Conditions

a. Data Splits and Preprocessing

The dataset is split into a training set with 11,550 samples and a test set with 2,888 samples. Each entry includes the full abstract (`medical_abstract`) and a categorical label (`condition_label`). The following preprocessing steps were performed:

- Renaming of the `condition_label` column to `label` to simplify model integration.
- Conversion of labels from the range [1–5] to [0–4] to ensure zero-based indexing.
- A label mapping dictionary was created to associate each numeric label with its corresponding textual class name
  (e.g., `0 → General Pathological Conditions`).
- Conversion of pandas DataFrames to Hugging Face `Dataset` objects to enable seamless integration with the `transformers` and `datasets` libraries.

These steps align with best practices in NLP preprocessing and help ensure consistency across model pipelines.

b. Class Imbalance

A crucial aspect of the dataset is its imbalanced class distribution, as visualized in *Figure 1* below. The category "General Pathological Conditions" dominates the dataset, followed by "Neoplasms" and "Cardiovascular Diseases". In contrast, "Digestive System Diseases" and "Nervous System Diseases" are underrepresented. This imbalance may bias model training toward majority classes and must be considered during evaluation.

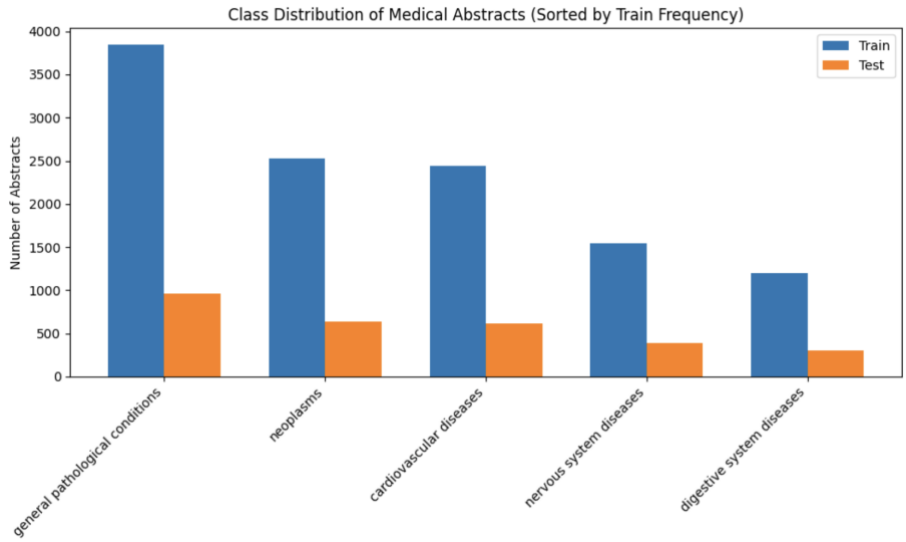

Figure 1: Class distribution in the training and test sets, sorted by training frequency.

3. Selected Approaches

The selection of models and techniques in this project was guided by domain relevance, interpretability, and empirical performance in biomedical NLP tasks. Specifically:

- *PubMedBERT* was chosen for its domain-specific pretraining on *PubMed* abstracts and full-texts, making it particularly suitable for medical text classification compared to general-purpose models such as *BERT* or *RoBERTa*. Prior studies have shown that domain-specific models significantly outperform general models in biomedical classification tasks (Gu et al., 2021).

- For embedding-based approaches, we selected pritamdeka/BioBERT-mnli-snli-scinli- scitail-mednli-stsb due to its fine-tuning on multiple biomedical (Lee et al., 2020) and natural language inference datasets, enabling it to capture semantic relationships relevant to medical abstracts. This model outperformed several alternatives in semantic similarity benchmarks on Hugging Face.

- Logistic Regression was used as the classifier for the embedding-based supervised approach due to its simplicity, interpretability, and robustness with small-to-moderate datasets. Alternative models such as SVMs or neural networks may offer higher accuracy, but at the cost of interpretability and increased training time—trade-offs that are less suitable in constrained environments.

- For clustering, UMAP (McInnes et al., 2018) was selected for dimensionality reduction because of its ability to preserve both local and global structures in high-dimensional semantic spaces, which is important when working with nuanced medical topics. K-Means was chosen due to its simplicity and compatibility with fixed cluster numbers (k=5), aligning with our known class distribution.

These choices reflect a balance between domain adaptation, computational efficiency, and methodological clarity. Future work could include ablation studies or head-to-head comparisons with alternative models such as SciBERT or ClinicalBERT to more precisely quantify trade-offs.

a. Classification using Representation Models (w/o fine-tuning)

In this approach, we utilize a pre-trained transformer-based language model as a fixed representation encoder to perform classification directly on the raw input abstracts. Specifically, the model used is PubMedBERT (`microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`), which has been trained on large-scale biomedical text corpora.

The classification is implemented using Hugging Face's high-level `pipeline` interface for text classification. Each abstract is tokenized with a maximum length of 512 tokens and passed through the model without modifying any of its internal weights. The model outputs a set of class probabilities for each input, and the predicted class is determined by taking the `argmax` over these scores.

This inference-only setup is computationally efficient, requires no training or gradient updates, and can serve as a strong zero-shot-style baseline. However, since the model has not been specifically fine-tuned

for the given classification labels or dataset structure, it may lack sensitivity to domain-specific patterns and struggles particularly with class imbalance and semantic overlap among disease categories.

Figure 2: Direct classification using a pre-trained representation model without weight updates.

b. Classification using Representation Models (w/ fine-tuning)

The second approach builds upon the same pre-trained PubMedBERT model but enhances it through task-specific fine-tuning on the labeled training data. In this supervised setting, the model parameters, including those of the transformer encoder, are updated via gradient descent to better align with the specific structure and vocabulary of the medical abstracts.

The input data is first tokenized into `input_ids` and `attention_mask` using the model's tokenizer. Subsequently, Hugging Face's `Trainer` API is used to manage the fine-tuning process. The training configuration consists of three epochs, a learning rate of 2e-5, a batch size of 16, and weight decay set to 0.01. These settings were chosen based on common defaults for transformer fine-tuning and adjusted to balance training efficiency with generalization performance.

Fine-tuning allows the model to internalize label-specific and task-dependent patterns, which typically leads to improved classification performance, especially in distinguishing semantically similar categories. However, this improvement comes at the cost of increased computational complexity, longer training times, and the potential risk of overfitting, particularly in cases where the dataset is small or imbalanced.

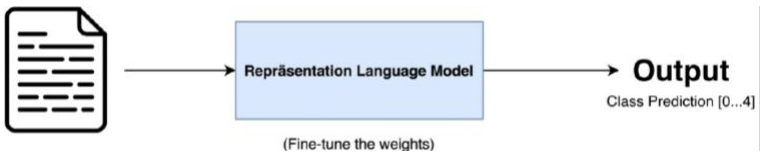

Figure 3: Fine-tuning a representation model by updating all parameters during supervised training.

c. Classification using Embeddings (Supervised Classification)

This approach follows a two-step architecture in which text embeddings are first generated using a pre-trained model, and a separate supervised classifier is trained on top of them. In our implementation, the SentenceTransformer model `"pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb"` was used to transform each abstract into a fixed-length embedding vector. This model has been fine-tuned on a mixture of biomedical and natural language inference datasets and is known to perform well in semantic similarity tasks within the medical domain.

In the first step, both the training and test abstracts were encoded into dense vector representations using the model's `.encode()` method with the parameter `convert_to_numpy=True`. This resulted in two matrices: one for the training set and one for the test set, where each row corresponds to a semantically meaningful embedding of a medical abstract.

In the second step, we trained a logistic regression classifier (`sklearn.linear_model.LogisticRegression`) on the training embeddings with their corresponding class labels. The classifier was then used to predict the labels of the test embeddings. The final output was compared against ground truth to compute evaluation metrics.

This architecture is lightweight and computationally efficient, as the transformer model is only used once to generate embeddings. It also allows easy substitution of the classifier, making it adaptable. However, because the embedding model remains frozen, it cannot adjust to task-specific idiosyncrasies, and the final performance relies on the expressiveness of the pre-trained embeddings.

Figure 4: Supervised classification pipeline using fixed embeddings and logistic regression.

d. Classification using Embeddings (zero-shot Classification)

The zero-shot classification approach implemented in this project also utilizes the `"pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb"` SentenceTransformer model, but in contrast to supervised learning, no classifier is trained. Instead, classification is performed by computing cosine similarity between the abstract embeddings and the embeddings of the class labels.

Concretely, we first embedded all test abstracts and all textual label names (e.g., *"neoplasms"*, *"nervous system diseases"*, etc.) using the `.encode()` method from SentenceTransformers. This yields two embedding spaces: one for the dataset samples and one for the labels.

A cosine similarity matrix was then computed using `sklearn.metrics.pairwise.cosine_similarity` between the test abstract embeddings and the label embeddings. For each abstract, the predicted class is the one with the maximum similarity score across the five label vectors.

This method is fully unsupervised and does not rely on any labeled training data, making it highly attractive for low-resource settings or rapid deployment scenarios. Nevertheless, its accuracy depends heavily on the semantic match between label descriptions and the embedding space. Additionally, subtle distinctions between medically related categories may not be fully captured by label embeddings alone.

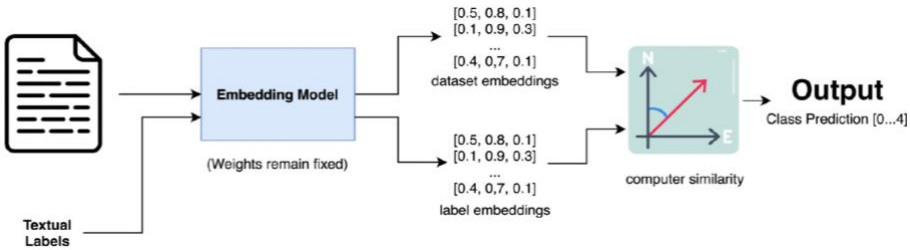

Figure 5: Zero-shot classification based on cosine similarity between abstract and label embeddings.

e. Classification using Text Clustering and Topic Modelling

The fifth approach explores an unsupervised learning strategy for classification by clustering medical abstracts in a lower-dimensional semantic space. The goal is to group similar abstracts without access to ground-truth labels during model training and to later assign cluster-based predictions using a k-nearest neighbor strategy.

The pipeline begins by encoding all abstracts using the same SentenceTransformer model as in previous sections: `"pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb"`. Each abstract is thus converted into a high-dimensional embedding vector that reflects its semantic content.

To make the clustering tractable and more interpretable, dimensionality reduction is applied using UMAP (Uniform Manifold Approximation and Projection), reducing the embedding space to five dimensions. UMAP preserves the local and global structure of the data, making it well-suited for biomedical text embeddings.

Subsequently, K-Means clustering is applied to the reduced training embeddings, dividing the dataset into five clusters—matching the number of known categories. Though unsupervised, these clusters are intended to correspond approximately to the disease categories based on semantic similarity.

To assign labels to the test set, the same embedding and dimensionality reduction steps are applied. Then, a k-Nearest Neighbors (KNN) classifier (with `k=5`) is trained on the reduced, clustered training data, and used to classify test abstracts based on proximity in the reduced embedding space.

This method requires no manual annotation or labeled training data and is therefore valuable in exploratory settings, such as topic modelling or literature organization. However, performance may be limited by the quality of the embeddings, cluster separability, and the assumption that K-Means can capture complex biomedical topic structures.

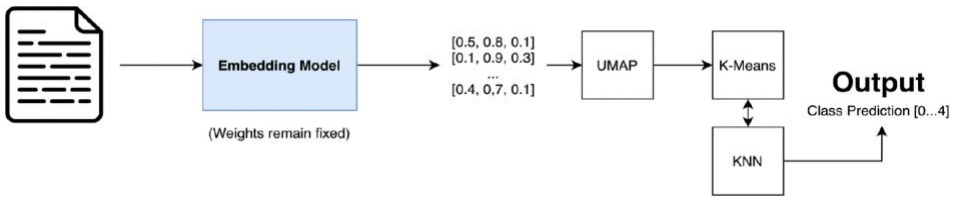

Figure 6: Clustering pipeline using UMAP for dimensionality reduction and KMeans + KNN for unsupervised classification.

4. Results

To assess the performance of the implemented classification approaches, we employed standard supervised learning metrics: accuracy, precision, recall, and F1-score. These metrics provide a multifaceted view of model quality, especially important in the context of imbalanced datasets, where accuracy alone can be misleading. In addition, we report macro-averaged and weighted-averaged scores.

All metrics were computed using `sklearn.metrics.classification_report`, and results were interpreted in the context of class distribution and model architecture.

a. Classification using Representation Models (w/o fine-tuning)

The results for the zero-shot pipeline using PubMedBERT without fine-tuning demonstrate that the model failed to generalize effectively to the task. As shown in *Figure 7*, the model predicted nearly all instances as "neoplasms", leading to perfect recall (1.00) for that class, but zero recall for all others.

The overall accuracy was 0.22, which corresponds roughly to the class frequency of "neoplasms" in the dataset, suggesting strong bias toward the majority class. The macro-averaged F1-score was only 0.07, highlighting the model's inability to distinguish between the five categories without task-specific supervision.

This outcome underlines the limitations of representation models used without fine-tuning in highly specialized, imbalanced medical classification tasks. Although PubMedBERT has been pretrained on biomedical corpora, its generic representations are insufficient for nuanced classification without adaptation.

	precision	recall	f1-score	support
neoplasms	0.22	1.00	0.36	633
digestive system diseases	0.00	0.00	0.00	299
nervous system diseases	0.00	0.00	0.00	385
cardiovascular diseases	0.00	0.00	0.00	610
general pathological conditions	0.00	0.00	0.00	961
accuracy			0.22	2888
macro avg	0.04	0.20	0.07	2888
weighted avg	0.05	0.22	0.08	2888

Figure 7: Evaluation metrics for representation-based classification without fine-tuning.

b. Classification using Representation Models (w/ fine-tuning)

Fine-tuning the same PubMedBERT model on the training set led to a substantial improvement in classification performance across all metrics. As shown in *Figure 8*, the model achieved a macro-averaged F1-score of 0.64 and an overall accuracy of 0.64.

All five classes saw considerable gains in both precision and recall, with F1-scores ranging from 0.49 to 0.75. The model also showed effective learning over three epochs, with training loss decreasing steadily, confirming convergence without severe overfitting.

These results validate the importance of task-specific fine-tuning for domain-sensitive classification. By allowing the model to update its internal weights, it adapts to label semantics and dataset distribution, even in the presence of imbalance.

Epoch	Training Loss	Validation Loss
1	0.928000	0.800302
2	0.756800	0.800431
3	0.623300	0.822237

	precision	recall	f1-score	support
neoplasms	0.69	0.81	0.75	633
digestive system diseases	0.53	0.66	0.59	299
nervous system diseases	0.59	0.66	0.62	385
cardiovascular diseases	0.68	0.82	0.74	610
general pathological conditions	0.63	0.41	0.49	961
accuracy			0.64	2888
macro avg	0.63	0.67	0.64	2888
weighted avg	0.64	0.64	0.63	2888

Figure 8: Evaluation metrics and training history for representation-based classification with fine-tuning.

c. Classification using Embeddings (Supervised Classification)

The supervised classification approach using static embeddings achieved solid and balanced results across most classes. As shown in *Figure 9*, the overall accuracy reached 0.57, with a macro-averaged F1-score of 0.57, indicating consistent behavior across all categories.

The performance for individual classes was relatively robust. The best-performing categories were *neoplasms* (F1 = 0.70) and *cardiovascular diseases* (F1 = 0.67), whereas *general pathological conditions* and *digestive system diseases* showed slightly lower F1-scores of 0.45 and 0.49, respectively. These results suggest that while the classifier was able to generalize reasonably well using the semantic embeddings, the separation between semantically similar classes remained a challenge.

Unlike transformer fine-tuning, this method did not update the embedding model. Thus, the performance relied heavily on the quality and discriminability of the pre-trained sentence embeddings. The logistic regression classifier showed good generalization, demonstrating the potential of hybrid architectures that combine deep representations with lightweight classifiers.

	precision	recall	f1-score	support
neoplasms	0.67	0.73	0.70	633
digestive system diseases	0.48	0.51	0.49	299
nervous system diseases	0.55	0.49	0.52	385
cardiovascular diseases	0.65	0.70	0.67	610
general pathological conditions	0.47	0.43	0.45	961
accuracy			0.57	2888
macro avg	0.56	0.57	0.57	2888
weighted avg	0.56	0.57	0.57	2888

Figure 9: Evaluation results for supervised classification using static sentence embeddings.

d. Classification using Embeddings (zero-shot Classification)

In contrast to the supervised version, the zero-shot embedding-based classification approach, shown in *Figure 10*, achieved slightly weaker results. The overall accuracy was 0.52, with a macro-averaged F1-score of 0.50. While these values remain acceptable given the complete absence of training, they fall short of the supervised alternatives.

The method performed best for the *cardiovascular diseases* and *neoplasms* categories, both achieving F1-scores above 0.64. However, recall dropped significantly for *general pathological conditions* (0.24) and *nervous system diseases* (0.34), which suggests that these classes were harder to match semantically using cosine similarity alone.

These findings confirm a key limitation of zero-shot approaches: the semantic alignment between abstract texts and short label descriptions is often insufficient to enable high-quality predictions. Nevertheless, for scenarios where training is not possible, this method provides a practical, lightweight baseline.

	precision	recall	f1-score	support
neoplasms	0.55	0.79	0.65	633
digestive system diseases	0.42	0.49	0.46	299
nervous system diseases	0.53	0.34	0.41	385
cardiovascular diseases	0.54	0.80	0.64	610
general pathological conditions	0.48	0.24	0.32	961
accuracy			0.52	2888
macro avg	0.50	0.53	0.50	2888
weighted avg	0.51	0.52	0.49	2888

Figure 10: Evaluation metrics for zero-shot classification using cosine similarity between embeddings.

e. Classification using Text Clustering and Topic Modelling

The clustering-based approach combines unsupervised representation learning, dimensionality reduction, and instance-based classification to assign labels to medical abstracts. It aims to uncover latent structure in the data without relying on explicitly annotated training samples.

The pipeline begins with embedding all abstracts using the `"pritamdeka/BioBERT- mnli-snli-scinli-scitail-mednli-stsb"` SentenceTransformer model. These high-dimensional vectors are then projected into a 5-dimensional latent space using UMAP, a non-linear dimensionality reduction technique that preserves both local and global relationships within the data.

Clustering is performed on the reduced embeddings using K-Means with `k=5`, aligning with the number of ground truth categories. Each abstract in the test set is processed through the same pipeline (embedding → UMAP) and classified using a K-Nearest Neighbors (KNN) model trained on the clustered training set.

The 2D UMAP visualization in *Figure 11* shows the spatial layout of embedded abstracts, colored by true class labels. The clusters overlap significantly, which indicates that the embedding space is not perfectly aligned with class boundaries, limiting separability under unsupervised assumptions.

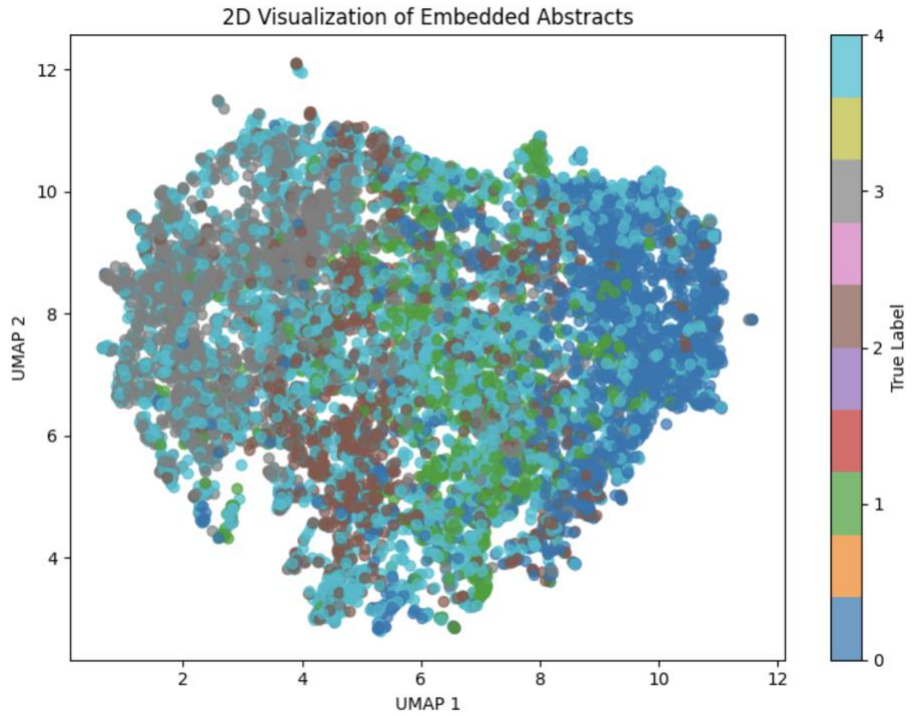

Figure 11: UMAP-based 2D visualization of test set embeddings colored by true class label.

Quantitatively, this approach achieved an overall accuracy of 0.52, with a macro-averaged F1-score of 0.51. Performance across categories varied, with *neoplasms* performing best (F1 = 0.68) and *digestive system diseases* the weakest (F1 = 0.40). These results are comparable to the zero-shot embedding method but clearly lower than those obtained via fine-tuned or supervised models.

	precision	recall	f1-score	support
neoplasms	0.62	0.74	0.68	633
digestive system diseases	0.37	0.43	0.40	299
nervous system diseases	0.50	0.47	0.49	385
cardiovascular diseases	0.55	0.65	0.60	610
general pathological conditions	0.45	0.33	0.38	961
accuracy			0.52	2888
macro avg	0.50	0.53	0.51	2888
weighted avg	0.51	0.52	0.51	2888

Figure 12: Evaluation metrics for clustering and KNN-based classification.

In conclusion, while this method is limited in accuracy, it offers useful exploratory insight and can be applied in scenarios where labeled data is unavailable or expensive to obtain. However, the results also highlight the difficulty of unsupervised biomedical text classification, particularly in semantically dense and overlapping domains.

5. Conclusion

In this study, we evaluated five transformer-based methods for classifying medical research abstracts into disease categories. The best results were obtained with the fine-tuned PubMedBERT model, which achieved a macro F1-score of 0.64 and accuracy of 0.64, demonstrating the effectiveness of task-specific adaptation in biomedical NLP.

The supervised embedding model also performed competitively (F1 = 0.57) while requiring significantly fewer computational resources. In contrast, zero-shot classification and clustering-based methods showed only moderate performance (F1 ≈ 0.50), highlighting the limitations of unsupervised or similarity-based techniques in fine-grained medical tasks.

The non-fine-tuned representation model performed poorly (F1 = 0.07), confirming that pretraining alone is not sufficient for precise classification without further adaptation.

Overall, the results confirm that fine-tuning remains essential for high-quality medical text classification, while embedding-based pipelines offer practical trade-offs in settings with limited training capacity.

6. Structure of Python Code

The structure of our code, along with some usage instructions, can be found in the README file.

7. References

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare, 3(1), 1-23.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 36(4), 1234-1240.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Journal of Open Source Software, 3(29), 861.

Sakai, H., & Lam, S. S. (2025). Large Language Models for Healthcare Text Classification: A Systematic Review. *arXiv preprint arXiv:2503.01159*.

Schopf, T., Onyema, C. O., & Ezenwoke, A. (2022). *Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches.* Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, and Humanities (LaTeCH-CLfL 2022).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J. (2020). *Transformers: State-of-the-Art Natural Language Processing.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.