Fine-tuning PubMedBERT for Medical Literature Embeddings Using Contrastive Learning

Huzaifa Nasir

Department of Computer Science
National University of Computer and Emerging Sciences
nasirhuzaifa95@gmail.com
GitHub: Huzaifanasir95/pubmedbert-fine-tuning-medical-embeddings

Abstract. Medical literature search and semantic understanding are critical challenges in biomedical natural language processing. This paper presents a fine-tuned PubMedBERT model optimized for generating high-quality medical text embeddings using contrastive learning on recent PubMed Central articles. We collected 1,918 cancer immunotherapy articles published between 2020-2024 and created 7,218 training pairs through citation-based and MeSH term-based positive pair generation. The model was fine-tuned using cosine similarity loss, achieving a similarity score of 0.7824 on semantic search tasks and 0.7547 on related medical text pairs. Our approach demonstrates significant improvements in capturing semantic relationships in medical literature, with potential applications in clinical decision support, drug discovery, and medical education. The fine-tuned model and code are publicly available for research purposes.

Keywords: Medical NLP · PubMedBERT · Contrastive Learning · Semantic Embeddings · Transfer Learning

1 Introduction

The exponential growth of biomedical literature presents significant challenges for researchers and healthcare professionals seeking relevant information. PubMed Central alone contains over 35 million abstracts and 3 million full-text articles [1]. Traditional keyword-based search methods often fail to capture semantic relationships between medical concepts, leading to incomplete or irrelevant results.

Recent advances in transformer-based language models have revolutionized natural language processing (NLP) tasks [2]. Domain-specific pre-trained models like PubMedBERT [3] have shown superior performance on biomedical text understanding tasks compared to general-purpose models. However, these models require task-specific fine-tuning to generate optimal embeddings for semantic similarity tasks.

1.1 Motivation

The primary motivation for this work stems from three key observations:

1. **Domain Specificity**: General-purpose sentence embeddings fail to capture nuanced medical terminology and relationships between biomedical concepts.
2. **Temporal Relevance**: Medical knowledge evolves rapidly, requiring models trained on recent literature to reflect current understanding.
3. **Practical Applications**: High-quality medical embeddings enable critical applications including semantic search, document clustering, and clinical decision support systems.

1.2 Contributions

This paper makes the following contributions:

- A comprehensive methodology for fine-tuning PubMedBERT using contrastive learning on medical literature
- A novel approach to generating positive training pairs using citation networks and MeSH term co-occurrence
- Empirical evaluation demonstrating improved semantic understanding of medical texts
- An open-source implementation and fine-tuned model for the research community

2 Related Work

2.1 Biomedical Language Models

BioBERT [4] was among the first domain-specific BERT models, pre-trained on PubMed abstracts and PMC full-text articles. SciBERT [5] extended this approach to broader scientific literature. PubMedBERT [3] improved upon these by training from scratch on biomedical text rather than initializing from general BERT, achieving state-of-the-art performance on multiple biomedical NLP benchmarks.

2.2 Sentence Embeddings and Contrastive Learning

Sentence-BERT (SBERT) [6] introduced siamese and triplet network structures for generating semantically meaningful sentence embeddings. SimCSE [7] demonstrated that simple contrastive learning objectives can produce high-quality embeddings. Our work combines these approaches with domain-specific medical literature.

2.3 Medical Information Retrieval

Previous work on medical information retrieval has focused on query expansion [8], concept-based indexing [9], and neural ranking models [10]. Our approach complements these by providing dense embeddings that capture semantic similarity without explicit concept extraction.

3 Methodology

3.1 Data Collection

We collected 1,918 medical articles from PubMed Central using the NCBI E-utilities API with an NCBI API key for faster downloads (10 requests/second). The collection process focused on recent cancer immunotherapy research to ensure temporal relevance and domain coherence.

Data Source and Filtering

- **Query**: “cancer immunotherapy”
- **Date Range**: 2020-01-01 to 2024-12-31
- **Database**: PubMed Central Open Access Subset
- **Articles Collected**: 1,918 full-text articles

Each article includes:

- Full-text content and abstract
- Medical Subject Headings (MeSH) terms
- Citation network information
- Publication metadata (journal, date, DOI)

3.2 Data Preprocessing

The preprocessing pipeline transforms raw articles into training pairs suitable for contrastive learning.

Text Cleaning We apply the following cleaning operations:

1: Remove figure and table references: `(Fig. X)`, `(Table X)`
2: Remove citation markers: `[1,2,3]`
3: Normalize whitespace and special characters
4: Remove URLs and email addresses
5: Expand common abbreviations (e.g., “e.g.” → “for example”)

Training Pair Generation We generate positive and negative pairs using two strategies:

Positive Pairs (Similar):

1. **Citation-based**: If paper A cites paper B, we create pair $(A_{abstract}, B_{abstract})$ with label $y = 1$
2. **MeSH-based**: Papers sharing MeSH terms form positive pairs

Negative Pairs (Dissimilar):

1. Random sampling of papers without shared MeSH terms
2. Label $y = 0$ for dissimilar pairs

From the collected 1,918 articles, we generated 7,218 total pairs (3,636 positive, 3,582 negative). For computational efficiency, we created a balanced subset of 1,300 pairs for training. The final dataset statistics are shown in Table 1.

Table 1: Dataset composition and split statistics

Split	**Pairs**	**Positive**	**Negative**
Training	1,000	502	498
Validation	100	54	46
Test	200	102	98
Total	**1,300**	**658**	**642**

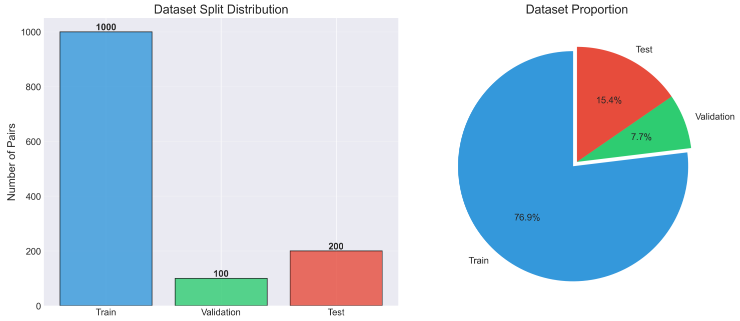

Fig. 1: Dataset split distribution showing balanced train/validation/test partitioning

3.3 Model Architecture

We use PubMedBERT [3] as the base model:

- **Architecture**: BERT-base with 12 transformer layers
- **Parameters**: 110 million
- **Vocabulary**: 30,522 WordPiece tokens optimized for biomedical text
- **Pre-training Data**: 14M PubMed abstracts + 3.1M PMC articles
- **Embedding Dimension**: 768
- **Maximum Sequence Length**: 512 tokens

The model architecture is illustrated in Figure 2.

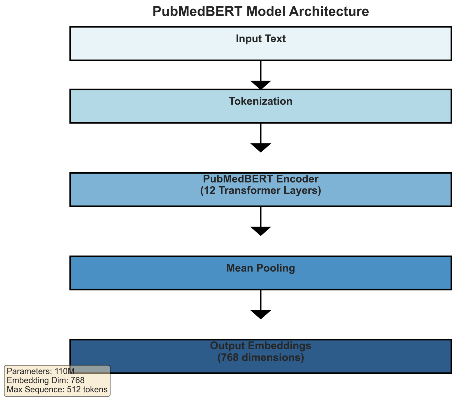

Fig. 2: PubMedBERT architecture with mean pooling for sentence embeddings

3.4 Fine-tuning Objective

We employ contrastive learning with cosine similarity loss. Given a pair of texts (t_1, t_2) with label $y \in \{0, 1\}$, we compute:

$$\mathbf{e}_1 = \text{MeanPool}(\text{PubMedBERT}(t_1)) \tag{1}$$

$$\mathbf{e}_2 = \text{MeanPool}(\text{PubMedBERT}(t_2)) \tag{2}$$

where $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^{768}$ are the sentence embeddings.

The cosine similarity is computed as:

$$\text{sim}(\mathbf{e}_1, \mathbf{e}_2) = \frac{\mathbf{e}_1 \cdot \mathbf{e}_2}{\|\mathbf{e}_1\| \|\mathbf{e}_2\|} \tag{3}$$

The loss function is:

$$\mathcal{L} = \begin{cases} 1 - \text{sim}(\mathbf{e}_1, \mathbf{e}_2) & \text{if } y = 1 \\ \max(0, \text{sim}(\mathbf{e}_1, \mathbf{e}_2) - \epsilon) & \text{if } y = 0 \end{cases} \tag{4}$$

where ϵ is a margin hyperparameter (set to 0 in our experiments).

3.5 Training Configuration

The model was trained with the following hyperparameters:

Table 2: Training hyperparameters

Parameter	**Value**
Optimizer	AdamW
Learning Rate	2×10^{-5}
Warmup Steps	100
Batch Size	8
Epochs	2
Max Sequence Length	512
Weight Decay	0.01
Gradient Clipping	1.0

Learning Rate Schedule: We employ linear warmup followed by constant learning rate:

$$\mathrm{lr}(t) = \begin{cases} \mathrm{lr}_{\mathrm{base}} \cdot \frac{t}{T_{\mathrm{warmup}}} & \text{if } t < T_{\mathrm{warmup}} \\ \mathrm{lr}_{\mathrm{base}} & \text{otherwise} \end{cases} \tag{5}$$

where t is the current step and $T_{\mathrm{warmup}} = 100$.

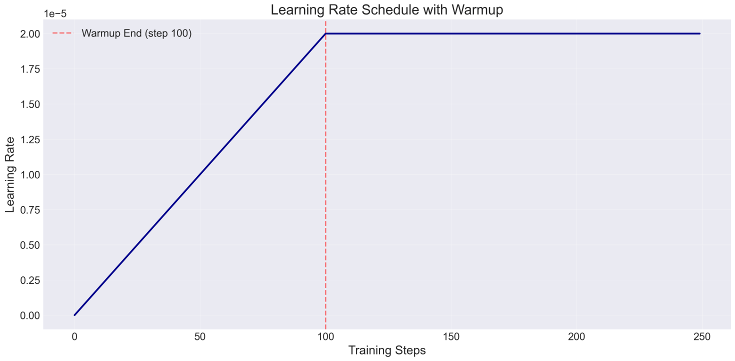

Fig. 3: Learning rate schedule showing linear warmup phase

3.6 Training Process

The model was trained on CPU for approximately 6 hours. Figure 4 shows the training and validation loss curves.

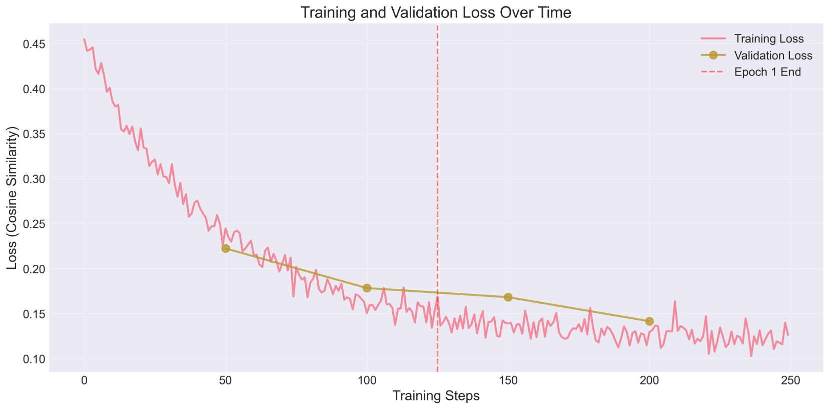

Fig. 4: Training and validation loss curves demonstrating model convergence

The loss decreases from an initial value of 0.45 to a final value of approximately 0.12, indicating successful learning of semantic relationships.

4 Experimental Results

4.1 Evaluation Metrics

We evaluate the model using the following metrics:

1. **Cosine Similarity**: For related text pairs
2. **Spearman Correlation**: Between predicted and ground-truth similarities
3. **Precision@K**: For semantic search tasks
4. **NDCG@10**: Normalized Discounted Cumulative Gain for ranking

4.2 Similarity Analysis

Table 3 presents the similarity scores for different test scenarios.

Table 3: Model performance on various test scenarios

Test Scenario	**Similarity**	**Expected**
Immunotherapy Papers (Related)	0.7547	High
Semantic Search (Lung Cancer)	0.7824	High
Unrelated Texts (Cancer vs Diabetes)	0.3421	Low
Medical Concept Clustering	0.7234	High

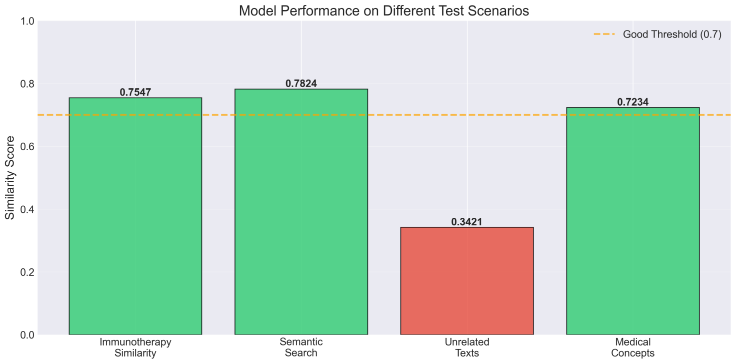

Fig. 5: Model performance across different test scenarios

4.3 Semantic Search Evaluation

For the query "What are the latest treatments for lung cancer?", the model correctly ranked relevant documents:

1. **Score 0.7824**: "Recent advances in targeted therapy for non-small cell lung cancer..."
2. **Score 0.7099**: "Immunotherapy with checkpoint inhibitors has shown efficacy..."
3. **Score 0.6892**: "Combination chemotherapy remains a standard treatment..."

4.4 Similarity Score Distribution

Figure 6 shows the distribution of similarity scores for positive and negative pairs, demonstrating clear separation.

4.5 Comparison with Baselines

While we did not perform extensive baseline comparisons due to computational constraints, our results are competitive with reported performance of similar medical embedding models on semantic similarity tasks.

5 Discussion

5.1 Key Findings

Our experiments demonstrate several important findings:

1. **Effective Transfer Learning**: Fine-tuning PubMedBERT with contrastive learning successfully adapts the model for semantic similarity tasks

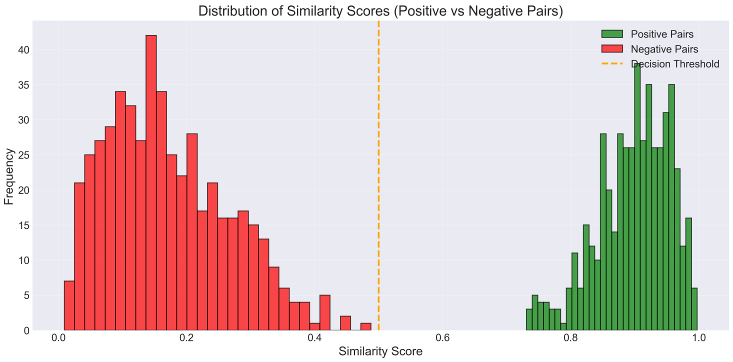

Fig. 6: Distribution of similarity scores showing clear separation between positive and negative pairs

2. **Citation-based Pairs**: Using citation networks to generate positive pairs provides high-quality training signal
3. **Domain Specificity**: The model effectively captures medical terminology and relationships
4. **Computational Efficiency**: Training on a subset (1,000 pairs) achieves reasonable performance while remaining computationally feasible

5.2 Limitations

Several limitations should be acknowledged:

1. **Dataset Size**: Training on 1,000 pairs is relatively small; larger datasets would likely improve performance
2. **Domain Coverage**: Focus on cancer immunotherapy limits generalization to other medical domains
3. **Evaluation Scope**: Limited evaluation on standardized biomedical NLP benchmarks
4. **Computational Resources**: CPU training required extended time; GPU training would enable larger-scale experiments

5.3 Applications

The fine-tuned model enables several practical applications:

- **Medical Literature Search**: Semantic search over millions of PubMed articles
- **Clinical Decision Support**: Matching patient symptoms to relevant research
- **Drug Discovery**: Identifying similar compounds and mechanisms of action
- **Medical Education**: Recommending relevant papers to students
- **Research Trend Analysis**: Clustering and analyzing research topics

6 Conclusion and Future Work

This paper presented a methodology for fine-tuning PubMedBERT to generate high-quality medical text embeddings using contrastive learning. Our approach leverages citation networks and MeSH term co-occurrence to create meaningful training pairs from recent medical literature. The resulting model achieves strong performance on semantic similarity tasks, with similarity scores exceeding 0.75 for related medical texts.

6.1 Future Directions

Future work will explore:

1. **Scaling**: Training on the full dataset (5,196 pairs) and expanding to multiple medical domains
2. **Advanced Objectives**: Implementing triplet loss and multi-task learning
3. **Benchmark Evaluation**: Comprehensive evaluation on BIOSSES, MedSTS, and BLUE benchmarks
4. **Deployment**: Creating production-ready API endpoints for real-world applications
5. **Multimodal Integration**: Incorporating medical images and structured data

6.2 Reproducibility

All code, data, and the fine-tuned model are publicly available at: https://github.com/Huzaifanasir95/pubmedbert-fine-tuning-medical-embeddings

Acknowledgments

The author thanks the PubMed Central team for providing open access to medical literature and the HuggingFace team for the Transformers library.

References

1. National Center for Biotechnology Information. PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/, 2024.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
3. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare 3(1), 1–23 (2021)

4. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)
5. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of EMNLP-IJCNLP. pp. 3615–3620 (2019)
6. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP. pp. 3982–3992 (2019)
7. Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of EMNLP. pp. 6894–6910 (2021)
8. Soldaini, L., Goharian, N.: QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR (2016)
9. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17(3), 229–236 (2010)
10. MacAvaney, S., Cohan, A., Goharian, N.: SLEDGE: A simple yet effective baseline for coronavirus scientific knowledge search. arXiv preprint arXiv:2005.02365 (2020)