Enhancing the Search for Scientific Articles: An Embedding Fine-Tuning Approach for Efficient Retrieval

Gabriel Ukstin Talasso, Ronaldinho Vega Centeno Olivera, Luis F. Solis Navarro, and Alexander Puma Pucho.

***Abstract*—Embedding models have significantly transformed text processing and natural language processing (NLP), enabling tasks like semantic representation and efficient information retrieval. However, the rapid expansion of scientific literature, particularly on platforms like arXiv, poses significant challenges for organizing and retrieving relevant documents efficiently. In this paper, we propose a two-stage fine-tuning process specifically designed to improve scientific literature retrieval. First, we apply adaptive pretraining using Masked Language Modeling (MLM) on the Arxiv corpus, allowing the model to better understand domain-specific language and structure. Subsequently, we perform supervised fine-tuning on the previously adapted model using contrastive learning to enhance its ability to retrieve relevant articles based on user queries. This approach trains the model to distinguish between relevant and irrelevant article-query pairs by maximizing the similarity between relevant pairs and minimizing it for irrelevant ones. Our experiments demonstrate that this two-step approach results in a more effective retrieval system, significantly improving the model's ability to match articles with user needs."**

***Index Terms*—Embedding models; Sentence embeddings; Scientific literature; Semantic representation; Information retrieval**

I. Introduction

Language embedding models have significantly transformed the fields of text processing and Natural Language Processing (NLP), achieving outstanding performance in tasks such as semantic text representation and information retrieval. Moreover, with the advent of pre-trained models like Word2Vec [1], BERT [2], and their variants, it has become easier to build systems capable of understanding complex linguistic structures. However, to fully exploit their potential in specific domains, an adaptation process is often required. This process tailors the model to a specialized dataset, enabling it to learn more precise representations for particular tasks.

Despite these advancements, the rapid expansion of scientific articles published on platforms like arXiv presents significant challenges in organizing and efficiently retrieving relevant documents. The growing volume of information makes it difficult for users to quickly find relevant research within specific thematic categories. Without an effective categorization system, document retrieval becomes slow and imprecise, hampering knowledge flow and evidence-based decision-making.

Therefore, this work aims to perform fine-tuning on a pre-trained embedding model, adapting it to the scientific literature domain. The fine-tuned model will generate dense vectors representing arXiv articles in a vector space, where similar articles are located in close proximity to one another. This adaptation involves training the model using domain-specific data from the arXiv corpus, followed by fine-tuning to optimize its performance in retrieving relevant documents. FAISS will be used to efficiently index and search these vectors, enabling fast and scalable retrieval of articles through similarity searches in an optimized search engine. This proposal seeks to improve the accessibility and organization of scientific information, providing researchers and users with faster and more accurate access to the most relevant documents in their areas of interest.

II. Objectives and Research Questions

A. Research Objective

The objective of this work is to adapt a pre-trained embedding model to the academic domain by fine-tuning it using a large corpus of scientific articles. The goal is to enhance the model's ability to generate more accurate semantic representations tailored to academic content, improving the efficiency and relevance of retrieving scientific papers. This adapted model will be integrated into a search system, aiming to improve the precision of results returned from keyword-based queries.

The main objective can be further broken down into two specific goals, which will guide the design and evaluation of the system:

- **Model Adaptation and Fine-tuning:** Investigate and implement strategies for fine-tuning a pre-trained model to better capture domain-specific knowledge within scientific literature. The aim is to enhance the model's performance by creating embeddings that are more relevant and accurate for academic research.
- **Creation of an Efficient Search System:** Design and develop an efficient search system for scientific articles that leverages the fine-tuned model. The goal is to optimize the article retrieval process, improving the accuracy

The authors are with the University of Campinas (UNICAMP), Institute of Computing (IC), Campinas, Brazil (e-mails: g2350785@dac.unicamp.br, r183585@dac.unicamp.br, l214616@dac.unicamp.br, a259936@dac.unicamp.br). All authors contributed equally to this work.

This paper was written for the MO810 course at the Institute of Computing, UNICAMP. The institute is located in Campinas, Brazil.

of the results and the relevance of the recommendations generated from scientific queries.

III. Methodology

The methodology consists of a two-stage approach aimed at adapting a pre-trained language model for the academic domain, specifically targeting the retrieval of relevant scientific papers. The first stage involves self-supervised pre-training of the model on a large corpus of scientific articles, while the second stage consists of supervised fine-tuning to optimize the model for semantic similarity-based retrieval tasks.

The overall workflow is outlined in Figure 1, illustrating the two main stages: domain-specific pre-training and supervised fine-tuning.

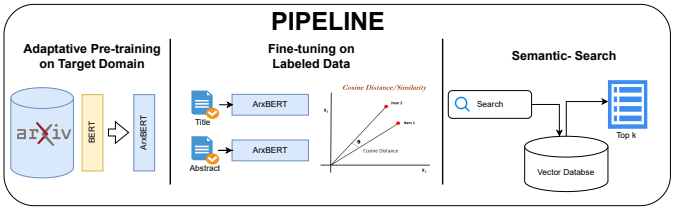

Fig. 1. Proposed Methodology Pipeline

A. Stage 1: Self-supervised Domain-adaptive Pre-training

In the first stage, we focus on adapting the pre-trained BERT model [3] to the scientific domain by continuing its training using a self-supervised approach [4]. Specifically, we utilize the **Masked Language Modeling (MLM)** task, where the model is trained to predict masked words in a sentence. This process enables the model to better capture the linguistic structures and terminology commonly found in scientific literature.

The model is pre-trained using a large corpus of scientific abstracts from the arXiv dataset [5], focusing on abstracts of research papers across various domains like physics, computer science, and mathematics. The pre-training is carried out without requiring labeled data, as MLM is a self-supervised task. See Figure 2.

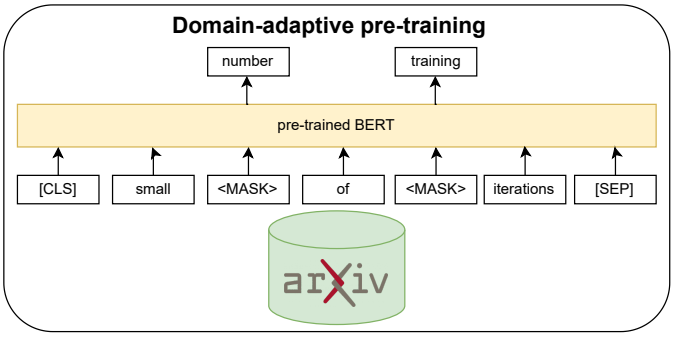

Fig. 2. Domain adaptation of BERT through MLM

B. Stage 2: Supervised Fine-tuning for Semantic Search

Once the domain-specific pre-training is complete, the next step is to fine-tune the model for the task of semantic search in scientific articles. The goal of this stage is to enable the model to generate embeddings that represent the semantic meaning of research articles, allowing for efficient similarity-based retrieval.

For this purpose, we employ a **Siamese Network** architecture [6], where two instances of the pre-trained model are used to process pairs of articles. The model is fine-tuned using a contrastive loss function that encourages the embeddings of semantically similar articles to be close together in vector space, while dissimilar articles are pushed farther apart, as shown in Figure 3. This enables more accurate similarity comparisons between user queries and the abstracts of scientific papers.

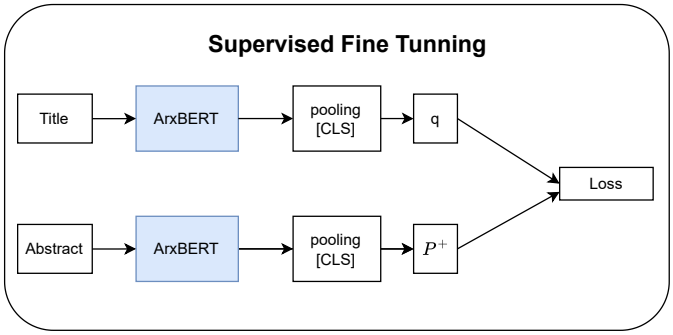

Fig. 3. Supervised fine-tuning of BERT using contrastive learning.

Loss Function: We use the **Multiple Negatives Ranking Loss** to fine-tune the model. This loss function is designed to optimize the model's ability to distinguish relevant (positive) pairs from irrelevant (negative) pairs. A positive pair consists of a query and its correct match, while all other samples in the batch are treated as negative pairs. One key feature of this loss function is the use of **in-batch negatives**, where all other examples in the same mini-batch serve as negative samples for each query. This approach makes training more efficient by using a large number of negative samples without the need to sample negatives externally.

The loss function is defined as follows:

$$\mathcal{L} = -\log\left(\frac{\exp(\text{sim}(q, p_+))}{\exp(\text{sim}(q, p_+)) + \sum_{i=1}^{N}\exp(\text{sim}(q, p_i))}\right)$$

Where $\text{sim}(q, p)$ represents the cosine similarity between the query q and the article p, and p_+ refers to the positive article (the correct match). The denominator sums the similarities of the query q with all other articles in the mini-batch, which are treated as negative samples (incorrect matches). The loss function encourages the model to rank the positive pair higher than any of the negative pairs in the batch, thus improving the model's ability to retrieve the most relevant articles.

Training Data Preparation: The training data for this fine-tuning process consists of pairs of articles from the arXiv dataset. Each pair is either a *positive pair*, where an article's title is paired with its abstract, or a *negative pair*, where the

abstract of an article is paired with the abstracts of other articles within the same mini-batch. The objective is to train the model to differentiate between semantically similar and dissimilar pairs.

- **Positive Pairs:** Each positive pair consists of an article's title q and its corresponding abstract p_+.
- **Negative Pairs:** Each negative pair consists of an article's abstract q paired with the abstract of another article p_i from the same mini-batch, where $i \neq +$.

C. Embedding Generation and Search Indexing

After the fine-tuning process, the model is used to generate embeddings for the abstracts of all articles in the dataset. These embeddings represent the semantic content of the articles in a high-dimensional vector space, where articles with similar content are located closer together.

To perform efficient and scalable similarity searches, we use **FAISS** (Facebook AI Similarity Search) [**?**], a library designed for fast similarity search in high-dimensional spaces. The FAISS index stores the article embeddings and allows for rapid retrieval of the most similar articles based on a user's query.

When a user submits a query, the query is first preprocessed and then converted into an embedding using the fine-tuned model. This embedding is used to search for the most similar articles in the FAISS index. The retrieved articles are ranked by their similarity to the query and presented to the user.

D. Evaluation Metrics

To assess the performance of the BERT model in searching for scientific articles, the **Accuracy@K** metrics (with $K = 1, 3, 5, 10$) were employed. These metrics are suitable for information retrieval tasks as they measure the model's ability to rank the correct abstract within the top K positions of the search results [7].

Accuracy@K is defined as the proportion of cases where the correct abstract appears within the top K positions of the list generated by the model:

$$\text{Accuracy@K} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(y_i \in \text{Top-K}\left(\hat{y}_i\right)\right)$$

where:

- N is the total number of evaluation samples.
- y_i is the correct abstract for the i-th title.
- $\hat{y}_i$ is the ranked list of abstracts retrieved by the model.
- $\mathbb{I}$ is the indicator function, which equals 1 if the correct abstract is within the top K positions and 0 otherwise.

Implementation of Metrics :For each article title in the test set, the BERT model generates a ranked list of abstracts based on predicted relevance. The metrics **Accuracy@1**, **Accuracy@3**, **Accuracy@5**, and **Accuracy@10** are calculated by checking if the correct abstract is within the top 1, 3, 5, and 10 positions, respectively:

- **Accuracy@1**: Proportion of titles where the correct abstract is the first result.
- **Accuracy@3**: Proportion of titles where the correct abstract is among the top three results.
- **Accuracy@5**: Proportion of titles where the correct abstract is among the top five results.
- **Accuracy@10**: Proportion of titles where the correct abstract is among the top ten results.

E. Datasets

The datasets used in this work are derived from the arXiv repository, which provides open access to scholarly articles across various scientific domains. We used two datasets:

- ***ArXiv Dataset*** This dataset contains over 1.7 million scholarly papers spanning multiple disciplines, including physics, computer science, mathematics, statistics, and more. Key features of the dataset include paper titles, authors, categories, abstracts, full-text PDFs, and additional metadata. This dataset is hosted on Kaggle and updated regularly. It is used as the base for the analysis and processing in this work [5].
- **CShorten/ML-ArXiv-Papers**This dataset is a filtered subset of the full arXiv dataset, specifically focusing on papers tagged with the "cs.LG" label, indicating they are related to Machine Learning. The dataset consists of approximately 100,000 papers and includes only the titles and abstracts of the papers, which are used for fine-tuning the embedding model. It is hosted on Huggingface and maintained through requests to the ArXiv API [1].

IV. Experiments

This section outlines the experimental setup and procedures undertaken to adapt and fine-tune the BERT model for the task of scientific article search. The experimentation was conducted in two main phases: domain adaptation of BERT for arXiv and supervised fine-tuning using Low-Rank Adaptation (LoRA) [8].

The code used to implement the experiments is publicly available in the GitHub repository [2]. This repository includes the scripts required to reproduce the results presented in this paper."

A. Domain Adaptation of BERT for arXiv

To tailor the BERT model to the specific domain of scientific literature available on arXiv, we performed domain adaptation through the following steps:

- **Model Freezing**: We froze 80% of the BERT model's parameters to retain the pre-trained knowledge while allowing the remaining 20% to be updated during training. This approach helps in preserving the general language understanding capabilities of BERT while adapting it to the specialized vocabulary and structure of scientific texts.
- **Training Data**: The adaptation was carried out using a dataset comprising 10,000 texts. Each text was constructed by concatenating the title of an article with its

[1]https://huggingface.co/datasets/CShorten/ML-ArXiv-Papers
[2]https://github.com/DinhoVCO/MO810-SSL/

corresponding abstract, providing the model with contextual information relevant to scientific discourse.

- **Masked Language Modeling (MLM)**: We employed a masked language modeling objective during training. This involved randomly masking certain words in the input texts and training the model to predict these masked tokens, thereby enhancing its ability to understand and generate scientific language.
- **Training Procedure**: The domain adaptation process was conducted over 10 epochs.

This domain adaptation phase aimed to align the BERT model more closely with the linguistic and structural characteristics of scientific articles, thereby improving its performance in downstream tasks such as abstract retrieval based on article titles.

B. Supervised Fine-Tuning with LoRA

Following domain adaptation, we performed supervised fine-tuning to further enhance the model's performance using Low-Rank Adaptation (LoRA). The specifics of this phase are as follows:

- **Low-Rank Adaptation (LoRA)**: LoRA was utilized to fine-tune the previously adapted BERT model. This technique involves introducing low-rank trainable matrices into each layer of the transformer architecture, allowing for efficient adaptation with a minimal increase in the number of parameters. Specifically, we set the rank parameter $r = 128$, which corresponds to approximately 4% of the total parameters of the original BERT model.
- **Training Data**: The supervised fine-tuning was performed on a dataset consisting of 2,000 pairs of titles and their corresponding abstracts.
- **Training Procedure**: The fine-tuning process was conducted over 5 epochs.

The supervised fine-tuning phase aimed to refine the model's ability to accurately associate article titles with their respective abstracts, thereby enhancing its retrieval performance in the context of scientific literature search.

C. Computing Resources

Both phases of the experimentation were conducted using the following resources:

- **Hardware**: The training processes were executed on a Colaboratory Notebook with an NVIDIA Tesla T4 GPU.
- **Software**: The models were implemented using Transformers, PyTorch, Sentence Tranformers and PEFT libraries.

V. Results and Discussions

To evaluate the performance of the proposed model, the CShorten/ML-ArXiv-Papers dataset from Huggingface was used. The metrics Accuracy@1, Accuracy@3, Accuracy@5, and Accuracy@10 were calculated. The results obtained are shown in Figure 4.

As shown in Table I, the Final Model (2 Stages) outperforms BERT Base in all the evaluated precision metrics. On the other hand, the Domain-Adapted model, which has only undergone the first stage of masked pretraining, performs worse than BERT Base across all metrics.

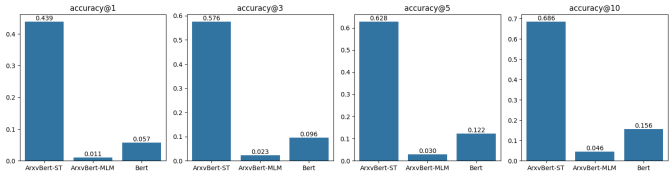

Fig. 4. The results of accuracy in the Top 1 (@1), Top 3 (@3), Top 5 (@5), and Top 10 (@10) ranges for the three evaluated models are presented in the image. These models include ArxvBert-ST, which combines ArxvBert-MLM with supervised fine-tuning; ArxvBert-MLM, a BERT-based model adapted to the specific domain; and BERT, a model pre-trained on general domains. The metrics show the performance of each model in correctly classifying examples within the first N positions of their respective predictions.

TABLE I
Precision Comparison Among Evaluated Models (Hypothetical Values)

Model	Accuracy@1	Accuracy@3	Accuracy@5	Accuracy@10
BERT-base	5.7%	9.6%	12.2%	15.6%
BERT-MLM	1.1%	2.3%	3%	4.6%
Final Model	43.9%	57.6%	62.8%	68.6%

The results obtained demonstrate that the proposed two-stage approach significantly enhances BERT's representational capabilities for the specific task of aligning titles with academic article summaries. The notable increase in precision metrics of the Final Model (2 Stages) compared to BERT Base suggests that supervised fine-tuning using the Multiple Negatives ranking loss function is highly effective in refining the model's semantic representations, improving the relevance and accuracy of the predictions.

On the other hand, the performance drop observed in the Domain-Adapted model indicates that masked pretraining alone is not sufficient to improve performance on the specific task of title-summary correlation.

VI. Conclusions and Future Work

Through adaptive pretraining and supervised fine-tuning, we have achieved a significant improvement in the BERT embedding model. The results obtained demonstrate that our model outperforms the base model, supporting the hypothesis that unsupervised pretraining alone is not sufficient to achieve optimal performance. In this regard, supervised fine-tuning with a small set of annotated data is a crucial step to enhance the model's semantic representation capabilities.

For future work, one potential extension would be to adapt the model to handle other types of academic texts, such as citations and abstracts, or even explore alignment tasks between other pairs of academic content, such as abstracts and keywords.

Another line of future research would involve evaluating the impact of incorporating a larger amount of training data in the first stage of pretraining, which could lead to further improvements in the model's performance on specific academic text correlation tasks.

Although **Accuracy@K** provides valuable insights into the model's performance, it does not fully capture the relative relevance of results beyond the presence of the correct abstract within the top K. Therefore, future work could incorporate additional metrics, such as **Mean Reciprocal Rank (MRR)** or **Mean Average Precision (MAP)**, to offer a more comprehensive assessment of the model's effectiveness.

VII. Acknowledgments

We thank Professor Dr. Marcelo da Silva Reis from the University of Campinas (UNICAMP), Institute of Computing (IC), for his valuable teaching in the subject of Self-supervised Learning, which served as the foundation for the development of this work.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

References

[1] T. Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, vol. 3781, 2013.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *CoRR*, vol. abs/2004.10964, 2020. [Online]. Available: https://arxiv.org/abs/2004.10964

[5] arXiv.org submitters, "arxiv dataset," 2024. [Online]. Available: https://www.kaggle.com/dsv/7548853

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[7] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, "Dense text retrieval based on pretrained language models: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2211.14876

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685