Information Retrieval System:

Domain-Specific Transformer Models for Binary Classification

Alvaro Gonzalez Mendez

Content

Assignment made by:

Álvaro González Méndez (190315) alvaro.gmendez@alumnos.upm.es

Introduction

Polyphenols are natural compounds synthesized by plants that have demonstrated significant relevance to human health due to their strong antioxidant activity. Given the vast volume of scientific research published in databases such as PubMed, finding relevant articles on this topic becomes a complex Information Retrieval (IR) task.

The objective of this work is the design, implementation, and evaluation of an IR system – specifically, a binary text classifier – capable of automatically identifying relevant scientific articles on polyphenol composition.

To address this problem, a comparative study of state-of-the-art (SOTA) methods was conducted. The methodology centers on the fine-tuning of pre-trained language models based on Transformers. Three architectures are evaluated: a generalist model (BERT) and two models specialized in the biomedical domain (BioBERT and BiomedBERT).

The models were trained using a corpus comprising 1,186 relevant and 1,185 non-relevant abstracts, collated specifically for this task. Subsequently, a comparative evaluation is conducted to select the optimal model, utilizing both classification and ranking metrics. Furthermore, a comparison with the results from the system developed by Cristina Rodriguez Fernandez and Didier Yamil Reyes Castro is included.

All development was conducted in Python utilizing Google Colab. The specific implementation details can be found in the Methodology section and in Appendix A. Code and Data.

State of the Art

Binary text classification is a fundamental task in IR. Traditionally, it was addressed using classic machine learning methods (such as SVM or Naive Bayes) applied to text representations like TF-IDF. However, in the last five years, the state-of-the-art (SOTA) has been completely dominated by Deep Learning models based on the Transformer architecture, proposed by Google in the paper "Attention Is All You Need" [1].

The turning point in the state-of-the-art (SOTA) of Natural Language Processing (NLP) occurred with the publication of BERT (Devlin et al., 2019). In their influential paper, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," the authors introduced a bidirectional Transformer architecture [2]. Unlike previous models, BERT is capable of reading an entire text sequence at once, understanding the context of a word based on the words that precede and follow it.

This model was pre-trained on a massive general-purpose text corpus (Wikipedia and BookCorpus), enabling it to learn deep and robust language representations, proving to be the preferred choice for a wide range of NLP tasks.

For the purposes of this project, the standard BERT implementation (specifically `bert-base-uncased`) represents the modern baseline. It is the fundamental generalist model against which all domain-specific models must be measured.

Following the success of generalist models like BERT [2], one of their key limitations was quickly identified: domain mismatch. It was demonstrated that a model pre-trained on general text (such as Wikipedia) does not possess the necessary lexical or semantic knowledge to understand highly specialized domains.

The scientific vocabulary, the syntax of abstracts, and the contextual meaning of words (e.g., acronyms) are drastically different from common language – a challenge that domain-specific models seek to resolve [3]. This need for specialization spurred the development of language models trained specifically for this domain, such as BioBERT [3] and BiomedBERT [4].

The first major advance in domain specialization was BioBERT [3]. In their paper, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Lee et al. (2020) proposed a domain adaptation method. Instead of training a model from scratch, they took the general BERT model [2] and continued its pre-training using millions of articles. The authors demonstrated that this continued pre-training significantly improved the model's performance on various biomedical text mining tasks, surpassing the generalist BERT.

For this project, BioBERT represents the first logical candidate for classification. Having been adapted to the biomedical corpus, it is expected to possess a better understanding of medical jargon and abstract semantics than the base BERT model.

The state-of-the-art evolved one step beyond simple adaptation. Researchers posited what would happen if, rather than adapting a general model, one was trained from scratch using exclusively domain-specific text. This approach materialized in the paper "Domain-specific language model pretraining for biomedical natural language processing" (Gu et al., 2020) [4]. This work introduced BiomedBERT (originally named PubMedBERT), a model trained solely on PubMed abstracts and full-text articles (PMC).

The authors demonstrated that this "pure" approach, which includes a completely domain-specific vocabulary and training regimen, outperformed the adapted BioBERT model [3] on most biomedical tasks. Given that the data for this project (both relevant and non-relevant) are drawn entirely from PubMed, BiomedBERT [4] is, theoretically, the SOTA model most aligned with the problem to be solved and the strongest candidate for the "best option."

Based on this review of the state-of-the-art, it is clear that the research favors domain-specific models [3][4] over general models [2] for biomedical tasks. However, the SOTA is not unanimous on whether domain adaptation (BioBERT) or training from scratch (BiomedBERT) is universally superior, as performance may depend on the specific task.

In this work, the goal is not only to build a classifier using the BERT [2] , BioBERT [3], and BiomedBERT (PubMedBERT) [4] models, but also to compare them and identify the best-performing of the three in order to achieve the best possible classification after fine-tuning all of them.

Methodology

This section discusses the procedural steps taken throughout the project. First, it addresses the extraction of polyphenol-relevant abstracts using the PubMed [6], Europe PMC [8], and Scopus [9] databases. Next, a corpus of non-relevant documents (i.e., those unrelated to these compounds) will be extracted, matching the size of the relevant corpus.

Subsequently, the three models – BERT, BioBERT, and BiomedBERT – will be fine-tuned on this complete dataset. Finally, their performance will be compared in the "Experimental Results" and "Discussion and Comparison" sections.

Data Collection and Corpus Creation

The teaching staff provided an Excel file containing metadata (title, authors, journal, publication date...) for 1,308 manuscripts, all related to polyphenols. The project's original aim was to retrieve their abstracts from PubMed using the NCBI E-utilities API [7]; however, only 663 of them were successfully obtained. Evidently, while this dataset might be sufficient for fine-tuning our models, the resulting models would not achieve the same quality as those trained on the complete corpus.

The second retrieval attempt involved using the Europe PMC article repository (as mentioned by a colleague in class), which yielded 962 abstracts. Subsequently, as the majority of the articles could be located manually in Google Scholar [10], the use of the Python library scholarly [5] was considered. This library performs web scraping to find information on the desired articles. Although this approach would theoretically solve the data scarcity problem, in practice, the database detects the execution of a script and proceeds to temporarily block the connection from the machine running it.

Finally, the chosen strategy was to first retrieve abstracts from Europe PMC and then search Scopus [9] for the 346 remaining documents. This process successfully retrieved 1,186 of the 1,308 documents. Although this does not constitute the complete corpus, it is considered sufficient to fine-tune the models and develop a robust classifier.

Furthermore, it should be noted that upon inspection of the provided Excel file, several entries were found that do not appear to correspond to any existing article (e.g., entry 1171, which is listed as ‘Not applicable’). These identified entries were removed during the creation of our dataset.

Once the corpus of 1,186 polyphenol-related abstracts was obtained, a second set of articles of a comparable size, which were not related to these compounds, was required. This was accomplished by utilizing the PubMed E-utilities API with the query 'NOT polyphenol'.

Finally, following these data retrieval operations, it was decided to merge both corpora (relevant and non-relevant) into a single dataset, saved as 'polyphenol_dataset.csv'. This dataset includes the article's title, its abstract, and a label indicating its relevance to polyphenols.

Preparing the data to train

To train our models, it is necessary to divide the dataset into three distinct subsets: a training set (encompassing 70% of the data) used for model fitting; a validation set (containing 15%), which allows us to observe the model's response to hyperparameter changes during the tuning process; and finally, a testing set (containing the remaining 15% of the articles), which is used to evaluate the efficacy of our models once they are fully tuned.

Lastly, it must be emphasized that it is crucial for all three sets to contain a balanced number of relevant and non-relevant samples. The cardinality of our data subsets, along with this aforementioned class balance, is reflected in Table 1.

	Number of Relevant Docs	**Number of Not Relevant Docs**
Training Set	830	829
Validation Set	178	178
Test Set	178	178

Table 1. Set sizes and distribution

Transformer Model Fine-Tuning

Once the data has been partitioned into the three subsets, the fine-tuning of BERT, BioBERT, and BiomedBERT can commence. For this purpose, the methodology followed the official Hugging Face tutorial on model training [11].

For the sake of simplicity, identical training arguments were used for all three classifier versions. Training was conducted for 10 epochs, with 208 steps per epoch, resulting in a total of 2,080 steps. The validation set was evaluated between epochs to monitor training performance and to save the model's state. This allows for checkpointing in case performance degrades.

The checkpoint (epoch) that yielded the highest accuracy on the validation set was selected as the final version for each model.

Given that the fine-tuning was executed in a free Google Colab environment using T4 GPUs, training each model took approximately 30 minutes. However, it is expected that training on modern, on-premise (non-cloud) hardware would result in significantly shorter training times.

To facilitate their use without requiring retraining, each fine-tuned model was saved in a separate zip file for straightforward loading.

Experimental Results

After training the three models, we can evaluate their performance using specific metrics. Two types of metrics can be applied: Set-Based and Rank-Based.

Set-Based Metrics

First are the set-based (cassification) metrics, which measure how well our model is able to differentiate between a relevant article and a non-relevant one. Measured on the test set and in the context of this project, the following metrics will be calculated for each model: Accuracy, Precision, Recall, and F1-Score. The results are presented in Table 2.

	BERT	**BioBERT**	**BiomedBERT**
Accuracy	0.9747	0.9887	0.9887
Precision	0.9751	0.9888	0.9887
Recall	0.9747	0.9887	0.9887
F1-Score	0.9747	0.9887	0.9887

Table 2. Set-based metrics for each of the three models

A brief inspection of Table 2 reveals that near-unity values were measured across all metrics. Typically, the closer a value approaches 1.0, the better the model is performing the task. Therefore, we can conclude that all three models fulfill the article classification objective exceptionally well.

During the calculation of the aforementioned metrics, the confusion matrices (Table 3, 4 and 5) for the test dataset were also generated.

	Pred Not Relevant	Pred Relevant
True Not Relevant	176	2
True Relevant	7	171

Table 3. BERT Confusion matrix

	Pred Not Relevant	Pred Relevant
True Not Relevant	175	3
True Relevant	1	177

Table 4. BioBERT Confusion matrix

	Pred Not Relevant	Pred Relevant
True Not Relevant	176	2
True Relevant	2	176

Table 5. BiomedBERT Confusion matrix

As can be clearly observed, the classifiers very rarely make classification errors. This is particularly true for the two specialized models, BioBERT and BiomedBERT, confirming that the use of domain-specific models does indeed enhance classifier performance. Although in our specific case the performance difference is small, the problem itself is relatively simple. It is plausible that this improvement would be more pronounced in more complex problems.

Rank-Based Metrics

Second are the ranking metrics, which measure the quality of an ordering over a set of documents. Although this type of metric may seem inapplicable to a classifier, it can be applied to the "logits" – the confidence score that indicates how 'certain' the model is of its classification.

Thus, by using this score to rank the articles from most relevant to least relevant, we can calculate the ROC Curve (Figure 4) and the Area Under this Curve (AUC-ROC) for each model. The AUC-ROC serves as a robust indicator of the quality of this ranking.

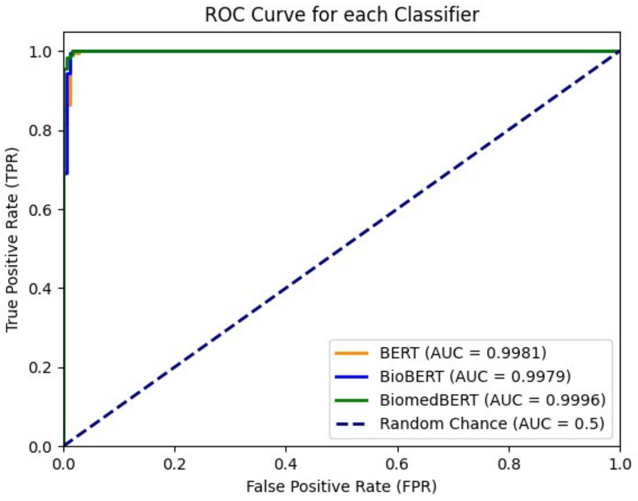

Figure 4. ROC Curve for each of the classifiers

Similarly to the classification metrics, we observe a very high AUC-ROC (approaching 1.0), which indicates that all classifiers are excellent and approach perfection (with respect to our test set). In particular, we can observe that the BiomedBERT model, which was trained exclusively on biomedical articles, obtains the highest score.

Other types of quality metrics exist, such as P@5, P@10, Reciprocal Rank (RR), R-Precision, and Average Precision. These have also been calculated, and their values for each model are shown in Table 6.

	BERT	**BioBERT**	**BiomedBERT**
AUROC	0.9981	0.9979	**0.9996**
P@5	1.0000	1.0000	1.0000
P@10	1.0000	1.0000	1.0000
RR	1.0000	1.0000	1.0000
R-Precision	0.9888	0.9888	0.9888
Avg. Precision	0.9980	0.9976	**0.9996**

Table 6. Rank-based metrics for each of the three models

As we previously observed with the set-based metrics, all three models demonstrate excellent performance. Upon also evaluating the rank-based metrics and inspecting Table 6, we obtained results indicating that the classifiers could also perform almost perfectly if they were utilized as rankers.

Discussion and Comparison

Now that all experiments have been conducted and all metrics for each model have been obtained, a comparison among the three models fine-tuned in this project will be performed.

Subsequently, the model considered to be the best-performing will be selected for a comparison against the classifier developed by another team in the course (Cristina Rodriguez Fernandez and Didier Yamil Reyes Castro).

Discussion of Results and Error Analysis

The experimental results demonstrate outstanding performance by all three implemented models. With F1-Scores exceeding 0.97 and AUROC values surpassing 0.997 in all cases, the Transformer architecture is confirmed as a good solution for this classification task.

As hypothesized in the State-of-the-Art review, the domain-specific models (BioBERT and BiomedBERT) outperformed the generalist model (BERT-base). Although the difference is minor, it is consistent: BioBERT and BiomedBERT achieved an F1-Score of 0.9888, superior to the 0.9747 achieved by BERT-base.

Analysis of the confusion matrices reveals the reason for this difference. The BERT-base model committed a total of 9 errors (7 False Negatives and 2 False Positives). In contrast, BioBERT and BiomedBERT each committed only 4 errors.

Deciding between BioBERT and BiomedBERT was challenging, as both achieved an identical F1-Score. To break the tie, it was necessary to examine their errors and ranking quality in greater depth.

In terms of errors, BioBERT performed slightly better by minimizing False Negatives (FN), committing only 1. This is arguably the most critical error in a retrieval system, as it conceals a relevant result from the user. BiomedBERT committed two FNs, which is still an excellent result compared to the 7 FNs from BERT-base.

However, where BiomedBERT truly excelled was in its ranking performance. It achieved an AUROC and an Average Precision of 0.9996, a near-perfect result. This indicates that it is exceptionally proficient at distinguishing between relevant and non-relevant documents.

Therefore, BiomedBERT was selected as the superior system. Although BioBERT had one fewer False Negative, BiomedBERT's superiority in global ranking metrics (AUROC and AvgPrec) demonstrates that its discriminative capability is the most robust.

Comparasion with other classmates' work

As the final part of this project, we were asked to optionally compare our information retrieval systems. Having previously collaborated with Didier and Cristina (who work as a team) on another assignment for this course, we decided to compare our respective classifiers.

It should be noted that, as I developed three distinct classifiers, I will only compare the one that, following the preceding analysis, yielded the best results (BiomedBERT).

The metrics used to compare the model proposed in this work and the one from my colleagues will be the same set-based and rank-based metrics previously used to ascertain the differences between the two projects. All these metric can be found on Table 7.

To complete their task, they also fine-tuned the pre-trained BioBERT model. Therefore, although I propose my fine-tuned BiomedBERT as the final classifier, the results of their work and mine should be very similar, as both models obtained comparable scores during my own project.

Upon inspecting Table 7, we can see that both models achieve very similar results, although the one proposed by my colleagues is marginally superior. This partially demonstrates that two analogous training procedures were conducted, and it is likely that the same classification results would be obtained for a given document when using either model.

However, a question remains to be resolved: Why does the information retrieval system proposed in this report perform worse than that of my colleagues?

Following a discussion with Cristina and Didier, we determined that we had used different methods for retrieving the abstracts from the web. This resulted in our datasets, while largely identical, being divergent. Their set contained a total corpus of 2,405 articles, which, compared to my 2,371, is slightly more complete.

We were training very similar models in a similar fashion, but their dataset was more complete. This discrepancy is the likely explanation for why my classifier obtained (slightly) worse results.

	BiomedBERT	**Colleagues Classificator**
	Set-Based Metrics	
Accuracy	0.9887	**0.9917**
Precision	0.9887	**0.9917**
Recall	0.9887	**0.9917**
F1-Score	0.9887	**0.9917**
	Rank-Based Metrics	
AUROC	0.9996	**0.9999**
P@5	1.0000	1.0000
P@10	1.0000	1.0000
RR	1.0000	1.0000
R-Precision	0.9888	**0.9917**
Avg. Precision	0.9996	**0.9999**

Table 7. Rank-based and Set-based metrics comparation with Didier and Cristina's work

Conclusions

Following the comparison with my colleagues' IR system, I would like to propose an idea that could potentially improve both projects. If the best components of both efforts were synthesized – specifically, by combining my colleagues' more comprehensive corpus with the use of a purely domain-specific pre-trained model (like BiomedBERT) – it is probable that even better results could be achieved.

As a final conclusion, this project has verified two key points: First, although generalist models like BERT perform well across a broad majority of fields, when undertaking tasks in domains with a specialized or highly specific lexicon, it is more effective to utilize models either adapted to that domain (BioBERT) or trained exclusively within it (BiomedBERT). Second, possessing a superior dataset, even by a marginal difference, can lead to superior results. Data must always be treated as the most critical component of the system.

On a personal level, this project served as my initial exposure to Transformer-based models, allowing me to learn how to fine-tune them, generate predictions, and measure their performance.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Kaiser, Ł. (2017). Attention is all you need. En *Advances in Neural Information Processing Systems* (pp. 5998-6008).

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171-4186).

[3] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

[4] Gu, Y., Tinn, R., Cheng, H., et al. (2020). Domain-specific language model pretraining for biomedical natural language processing. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2305-2316).

[5] Ibraim, G. I. G., et al. (2020). *scholarly: Scrape Google Scholar* [Software]. https://github.com/scholarly-python-package/scholarly

[6] U.S. National Library of Medicine. (n.d.). *PubMed* [Database]. National Institutes of Health. Retrieved November 11, 2025, from https://pubmed.ncbi.nlm.nih.gov/

[7] National Center for Biotechnology Information. (n.d.). *Entrez Programming Utilities (E-utilities)* [Software]. U.S. National Library of Medicine. Retrieved November 11, 2025, from https://www.ncbi.nlm.nih.gov/books/NBK25501/

[8] Europe PMC. (n.d.). *Europe PMC* [Database]. EMBL's European Bioinformatics Institute (EMBL-EBI). Retrieved November 11, 2025, from https://europepmc.org

[9] Elsevier. (n.d.). *Scopus* [Database]. Retrieved November 11, 2025, from https://www.scopus.com

[10] Google. (n.d.). *Google Scholar* [Database]. Retrieved November 11, 2025, from https://scholar.google.com/

[11] Hugging Face. (n.d.). *Transformers Documentation - Training*. Retrieved November 11, 2025, from https://huggingface.co/docs/transformers/training

Appendix A. Code and Data

For the completion of this project, the materials provided by the teaching staff, ‘publications.xlsx’, were used as the starting point.

Data processing, model training, and the evaluation of said models were all performed using Google Colab (free version) as the execution environment. The fitted models and the pre-processed data are located in the .zip file provided with this submission. They can also be found on GitHub at:

https://github.com/alvarogmendez/biomedical_info_retrieval