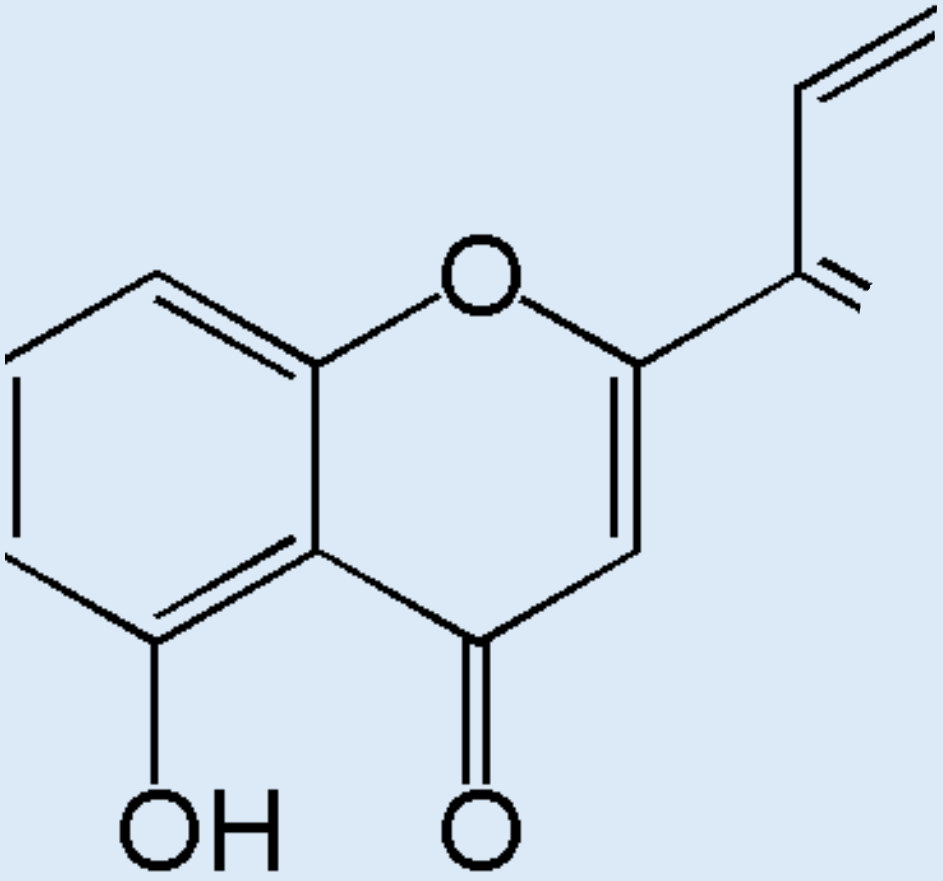
Information Retrieval System: Domain-Specific Transformer Model for Binary Classification

Alvaro Gonzalez Mendez

Content

- Assignment Problem and Goals
- Methodology
- Results
- Results Comparasion
- Conclusion



The problem

- Build an IR System that decides if an article talks about polyphenols.
- Possible solutions:
  - Classic Classifiers
    - SVM
    - Logistic Regression
    - NaiveBayes
  - Modern Technologies
    - Transformers

Models to be used

Fine-tuning pre-trained Language Models.

I've trained:

- Generalist → BERT
- Adapted Model (Domain-Adapted) → BioBERT
- From scratch Model (Domain-Specific) → BiomedBERT

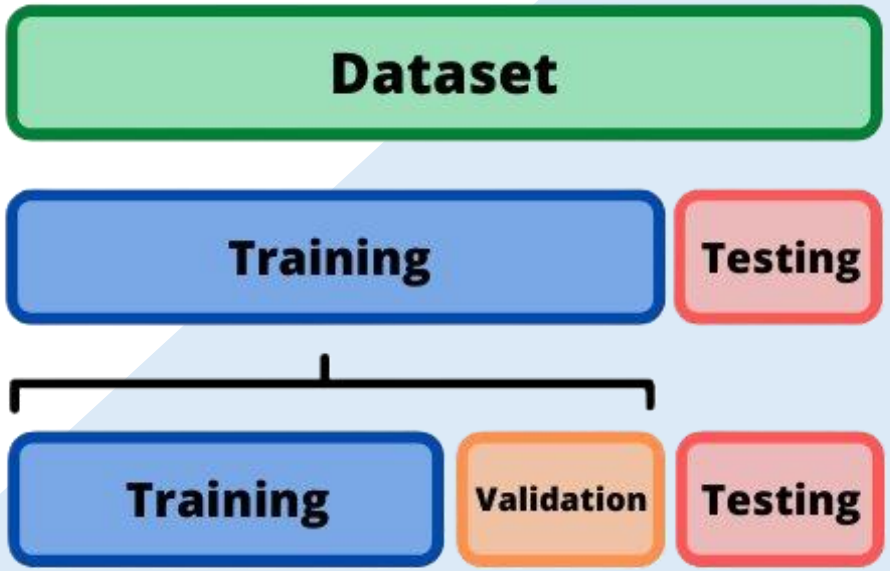
Gu, Y., Tinn, R., Cheng, H., et al. (2020). Domain-specific language model pretraining for biomedical natural language processing

Methodology: Data Retrieval

- From the 1308 relevant papers → retrieved 1186 abstracts
- Searched relevant abstracts in EuropePMC and Elsevier Scopus
- Searched not relevant on PubMed
- Balance between relevant and not relevant

Methodology: Experimental Setup

- 1. Divide our dataset in 3:
  - 70% → Train set
  - 15% → Validation set
  - 15% → Test set
- 2. Train the 3 models.
- 3. Test the models with the test set and get the metrics.
- 4. Find the most outstanding one



Results: Classification Metrics

BERT

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

BioBERT

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

BiomedBERT

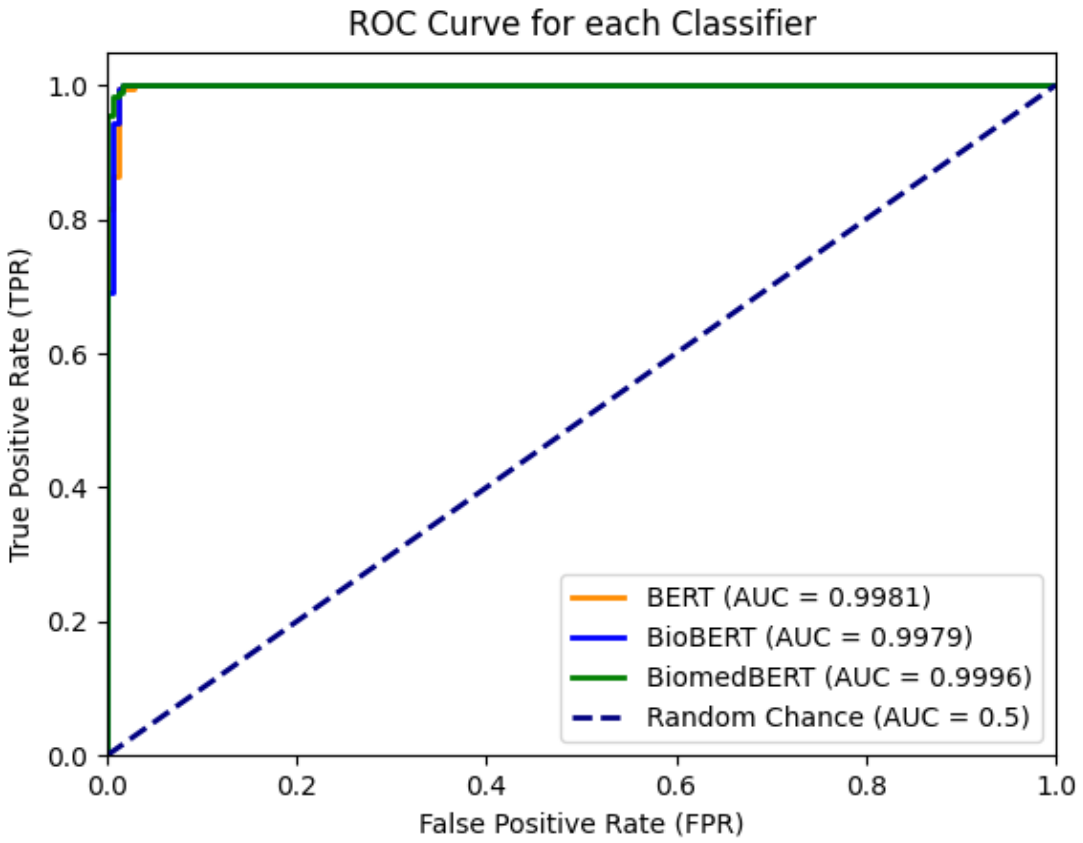
	Pred Not Relevant	Pred Relevant
True Not Relevant	176	2
True Relevant	2	176

Total of 356 docs on the test set were used to obtain these confusion matrix

Results: Classification Metrics

	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

Results: Ranking Metrics



- How to measure ranking on a classifier?
- Ranking the classification results using the logits.
- Logits → How confident is the model with the classification of an element.

Results: Ranking Metrics

	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**

BERT vs BioBERT vs BiomedBERT

Generalist model does good but worse than the specific ones.

Models trained with a biomedical domain tied in most of the classification metrics.

BiomedBERT performs slightly better in ranking metrics.

- Higher AUROC
- Higher Avg Precision

Comparison with colleagues

Compared all the set-based and rank-based measures with Didier and Cristina.

They have used BioBERT too but obtained slightly better results.

Most likely to happened because the had made a bigger and more complete dataset by 1200 relevant documents

	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**

Conclusions

After training and comparing all three models + comparing with the colleagues work I built my own conclusions:

- Specific-Domain and Adapted-Domain transformers work better than general.
- Obtained an almost perfect classifier.
- Data size matters. Is as important as choose the model to train.

Posible tweak for getting better experiment results → Using Cristina and Didier's dataset to fine-tune BiomedBERT

Thank You!

Any Questions?