ArcTEX – a precise clinical data enrichment model to support real world evidence studies

Joseph Cronin, PhD, Keiran Tait, Jamie Wallis, PhD, Robert Dürichen, PhD
Arcturis Data, Building One, Oxford Technology Park, Technology Drive, Kidlington, OX5 1GN UK

Introduction

Leading pharmaceutical companies use real-world data, requiring specific biomarker values. These values are frequently found in unstructured text, hindering direct analysis. Privacy concerns and resource constraints also limit in-hospital analysis. To address this, we developed ArcTEX (Arcturis Text Enrichment and EXtraction), a lightweight QA model that accurately extracts biomarker information from unstructured clinical reports. ArcTEX is flexible, requiring few training samples to adapt to other biomarkers, robust with confidence scores to identify misclassified samples, and CPU-executable, ensuring patient data privacy for hospital use.

Objectives

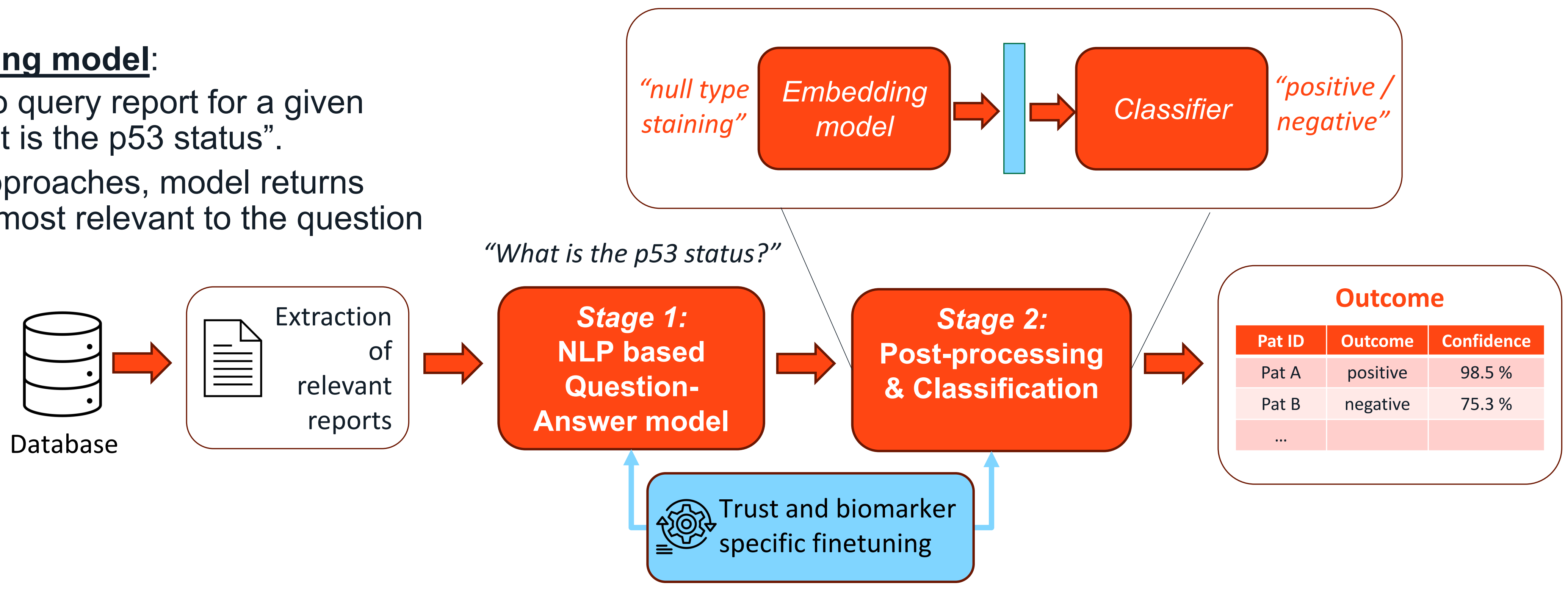
- Develop a generic approach to extract **biomarker values** and **supplementary disease information** from free text reports across trusts to support RWE demand
- Focus on **high precision** to support ongoing RWE studies (incl. confidence scores)
- Develop framework to monitor and further improve performance of algorithms to increase trust of customers and support future regulatory submissions
- Ensure that computation can be **performed on local/on-prem infrastructure** (e.g. data extraction on trust side)
- Ensure the approach does not return personal identifiable information to increase acceptance at trust side

Methods

2-stage process: QA stage and classification stage

STAGE1 - Question-Answering model:

- BioBERT QA model used to query report for a given biomarker status, e.g. "what is the p53 status".
- Compared to generative approaches, model returns subsection of original text most relevant to the question (no hallucinations).


"null type staining"
Embedding model
Classifier
"positive / negative"
"What is the p53 status?"
Database
Extraction of relevant reports
Stage 1: NLP based Question-Answer model
Stage 2: Post-processing & Classification
Outcome
Pat ID | Outcome | Confidence
Pat A | positive | 98.5 %
Pat B | negative | 75.3 %
…
Trust and biomarker specific finetuning

STAGE2 – classification model:

- Use sentence embedding model with classification head to output the predicted biomarker class, e.g. positive/negative.
- SetFit[2] approach requires a small amount of training data – one label file per biomarker/question that links answers to positive/negative classes
- TSDAE[3] unsupervised pretraining utilized to improve model accuracy.
- Model pretrained on sentences from entire dataset.

Datasets:

- 77,693 anonymised English pathology reports OUH.
- Patients have at least one of the 7 oncology areas: lung, pancreatic, renal, breast, ovarian, endometrial, or liver.
- The length of the reports varies between 20-4015 characters (average: 1084).
- **Finetuning dataset**: subset of 243 annotated reports for 14 biomarkers (e.g. p53, er, pr, her2, mmr, tumour grade).
- **Extended validation & test datasets (EVT)**: additional 200 annotated reports for each biomarker

Evaluation Schema:

- 50 reports randomly selected from the EVT dataset to make up both a test and validation set.
- Model evaluated on the validation set. At each iteration, the worst 5 classified samples (either misclassified and/or samples with the lowest confidence score) were added to a training set in the next iteration (the validation set was replenished by 5 randomly selected examples of the remaining EVT dataset).
- Procedure repeated for 5 iterations, adding up to 25 additional training samples.
- Evaluation repeated 10 times to estimate the robustness of the approach.
- **ArcTEX model:** finetuned BioBERT[1] model in stage 1 and a domain adapted setfit classifier in stage 2.
- **Baseline models**: BERT QA models in stage 1 and a non-domain adapted setfit classifier in stage 2.

Results

Results in Figure 1 indicate that mean accuracy for the different non-finetuned, non-domain adapted BERT models is between 84.8-88% at iteration 0. The avg. accuracy can be increased by finetuning of a BioBERT model to 92.2% and by performing additional domain adaptation (ArcTEX model) to 93.6%. Compared to the baseline models, adding additional challenging training samples only improves the performance slightly, indicating that little to no further training data is required. This is confirmed by Table 1 for other biomarkers.

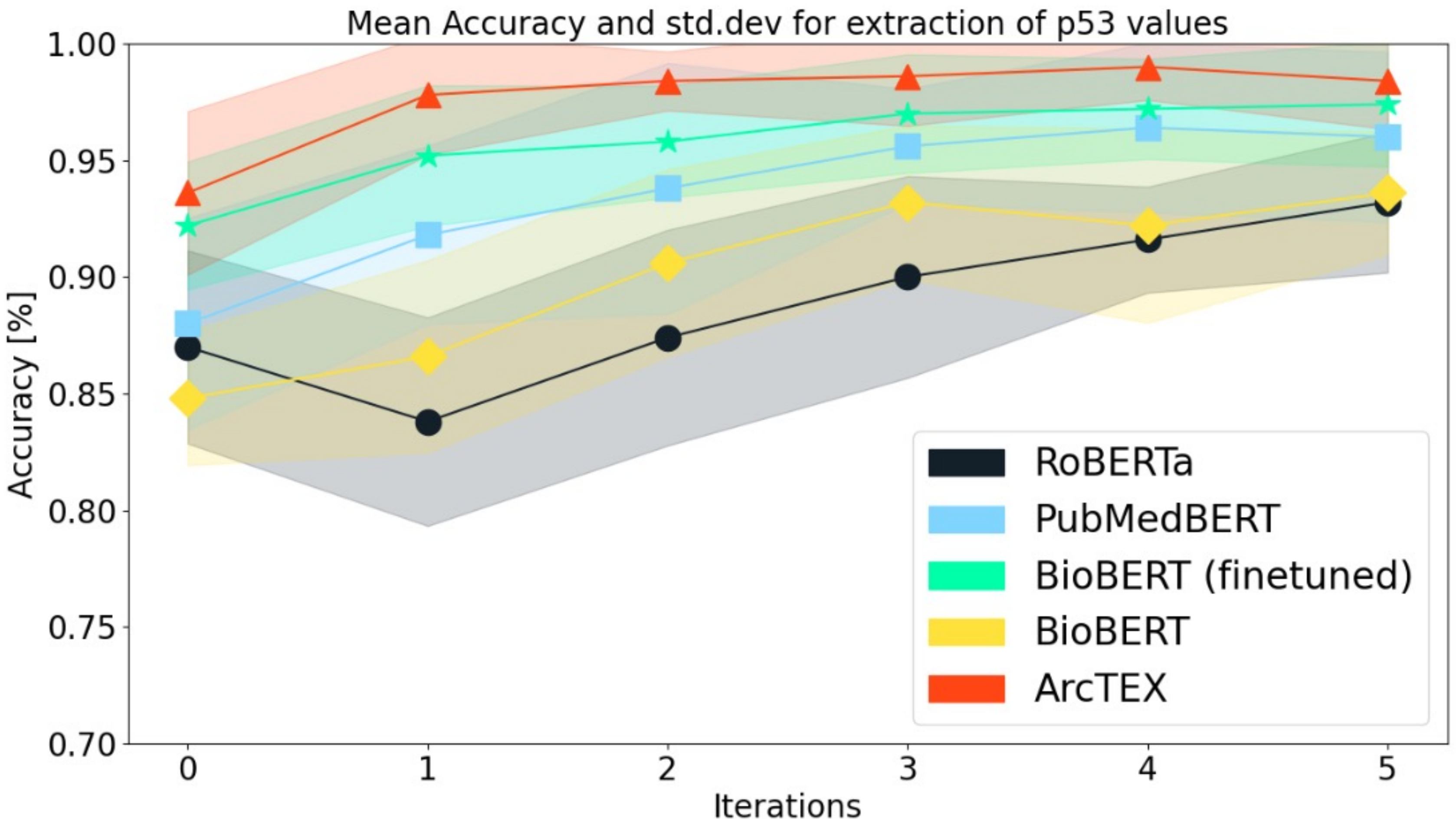


Figure 1 – Mean accuracy and standard deviation for different QA models to extract the status of biomarker p53 from unstructured reports.

Table 1 - mean accuracy (standard deviation) for BioBERT and ArcTEX model for different biomarkers at iteration 0 and 5.

Marker	BioBERT		ArcTEX	
	iter. 0	iter. 5	iter. 0	iter. 5
P53	84.8 (2.9)	93.6 (2.6)	93.6 (3.5)	**98.4 (2.1)**
MSH6	92.8 (3.4)	96.5 (2.1)	98.2 (2.0)	**99.0 (2.5)**
Grade	69.4 (7.9)	98.0 (1.6)	92.6 (3.0)	**98.2 (2.0)**
FIGO	89.8 (3.3)	98.6 (1.3)	95.6 (2.8)	**99.0 (1.1)**
MMR	81.6 (5.6)	95.4 (5.2)	96.6 (3.8)	**99.6(0.84)**

Conclusions

- We demonstrate that through unsupervised domain adaption, intelligent classification (setfit) and finetuning, the ArcTEX model can extract biomarker values and disease relevant information with high accuracy and little training samples
- This approach does not rely on large language models and through the integrated classification stage it ensures that no personal identifiable information is released, making it suitable for hospital environments.
- In the next stage, the model will be evaluated in a hospital environment and further validated by clinical experts.

Arcturis

References

1. J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/BIOINFORMATICS/BTZ682.
2. L. Tunstall et al., "Efficient Few-Shot Learning Without Prompts," Sep. 2022, Available: http://arxiv.org/abs/2209.11055
3. K. Wang, N. Reimers, and I. Gurevych, "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning," Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, pp. 671–688, Apr. 2021, doi: 10.18653/v1/2021.findings-emnlp.59.

Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support. We believe using patient data is vital to improve health and care for everyone and would, thus, like to thank all those involved for their contribution. A pseudonymised dataset for this study was provided by the Oxford University Hospital Trust.