UNIVERSITY OF LEEDS

Evaluating Text Pre-Processing Strategies for Clinical Document Classification with BERT

Authors: Sarah Miller[1], Serge Sharoff[2], Geoffrey Hall[2], & Prabhu Arumugam[3]

Presented by: Sarah Miller, UKRI CDT in AI for Medical Diagnosis and Care, School of Computing, University of Leeds. Email: scslmi@leeds.ac.uk

IARIA

UNIVERSITY OF LEEDS

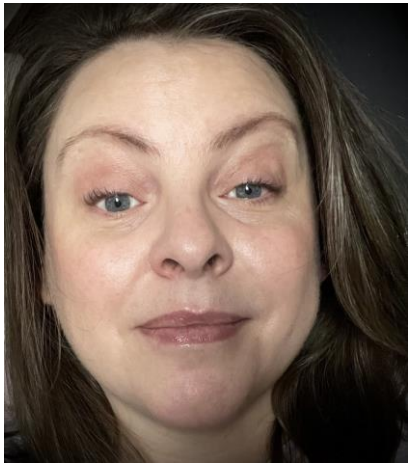
Biography

Sarah Miller is a doctoral student in the UKRI CDT in AI for Medical Diagnosis and Care, School of Computing, University of Leeds.

Research: her research interests are the application and suitability of pretrained language models for clinical NLP tasks.

Education: prior to her PhD she obtained masters degrees in Artificial Intelligence and Data Science, and an undergraduate degree in computing.

Industrial Experience: She worked as a Business Intelligence Developer for the UK's largest tour operator.



Introduction

- Extracting information from clinical texts is currently a manual task for the clinically trained and it is both time consuming and costly for healthcare providers.
- Automating this task with Natural Language Processing (NLP) has the potential to deliver efficiencies, saving both time and money.
- Bidirectional Encoder Representations from Transformers (BERT) models have delivered notable results in many NLP tasks.
- But… adopting these models for use with clinical documents comes with challenges.

Introduction

- BERT models have limitations for the size of texts sequences they can accept as input.
- BERT models accept only 512 tokens, and tokens are not equivalent to words. Tokens are word pieces and clinical texts often exceed the maximum limit.
- A solution to handle longer texts is pre-processing them to accommodate the text size limit.
- But.. clinical documents are variable in length and structure making them difficult to process.

Aims of the study

- In this study we aimed to investigate the challenges of applying BERT models to a clinical document classification task.
- To investigate how various methods of text pre-processing impacted document classification results - over varying document lengths.
- To understand how different variants of BERT models handled the task.

Contributions

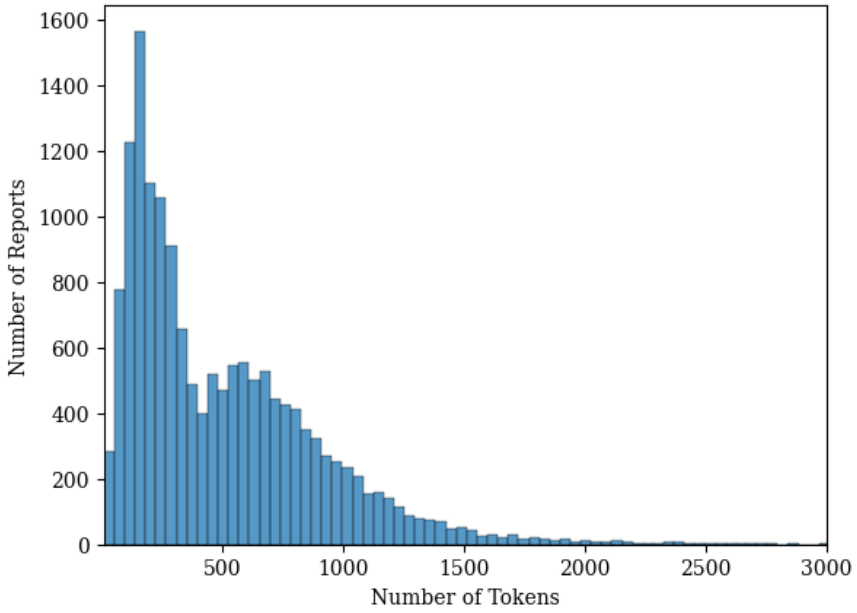
- Our experiments are performed with a novel dataset for this task – the limited studies in this area predominately are carried out using the MIMIC-III discharge summaries for classifying ICD9 codes.
- To the best of our knowledge is the only study in this area that examines a wider variation of text pre-processing methods and multiple variants of BERT models to pathology report texts.
- To the best of our knowledge is the only study to examine performance over variations of text length distributions.

UNIVERSITY OF LEEDS

Methodology - Dataset

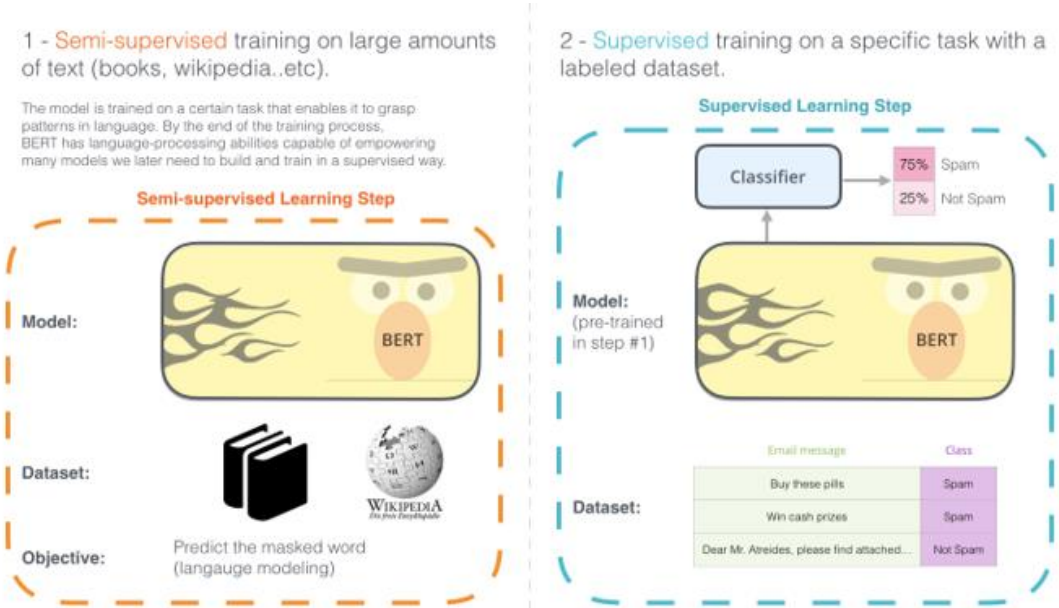
Column Label	Dataset Distribution per label/class label		
	Class labels	Total per Class label	Total No. Reports
Disease Type	Breast	7767	15825
	Colorectal	6389	
	Lung	1668	
Histology ICD-0-3 Code	80703	985	15825
	81403	6664	
	84803	628	
	85003	6310	
	85203	1238	
Grade	G1	878	15825
	G2	9647	
	G3	4436	
	G4	<5	
	GX	861	

Pathology report texts from Genomics England. The reports are variable length, shortest just 10 tokens and the longest 5372. Mean token length is 501, with 25% of the reports exceeding 700 tokens and 25% of them being less than 200 tokens.



Methodology – Models & Hyperparameters

- BERT-base-uncased
- Bio_ClinicalBERT – trained on all PubMed (abstracts + full text) and all MIMIC-III texts.
- BiomedBERT – trained on all of PubMed (abstracts + full text)
- Trained for 3 epochs, with a batch size of16 and a learning rate of 3-5e

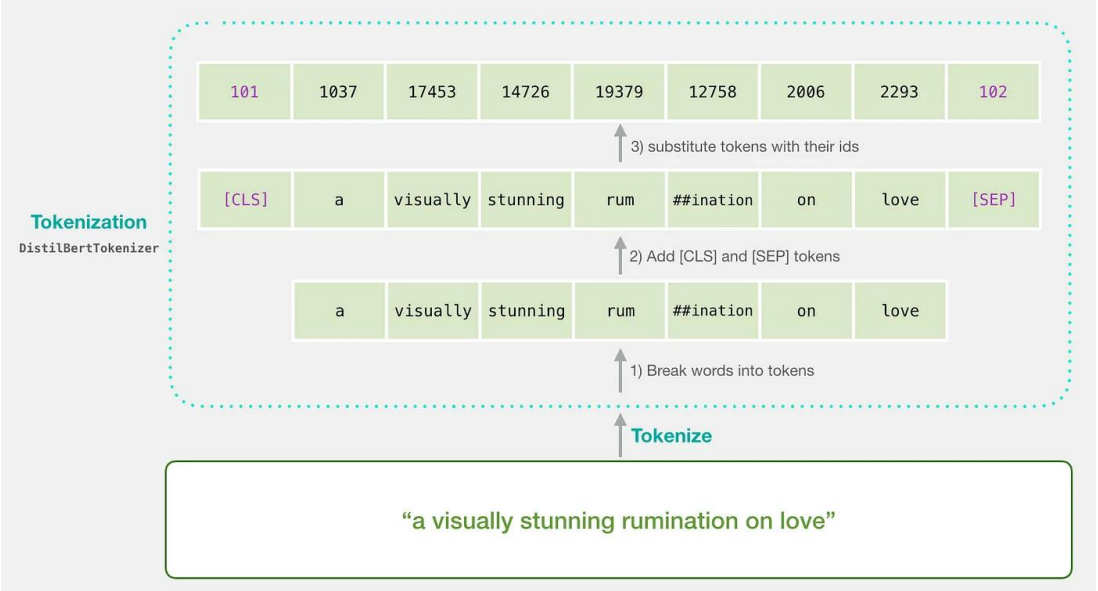


The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. via. (https://jalammar.github.io/illustrated-bert/)

UNIVERSITY OF LEEDS

Methodology – BERT Tokenisation

- The BERT tokenizer converts text sequences into word piece tokens. Word piece tokens are words that have been split into segments and transformed into numerical representations the model accepts as input.
- The word to token ratio given throughout literature is approx. 400 words = 512 tokens and because the word to token limit can only be approximated, we split documents using the token length.



BERT tokenisation process taken from: https://satish1v.medium.com/tokenization-for-bert-models-5c20734d1aca

Methodology – Text Pre-processing Strategies

- **Right and Left Truncation:** Truncation of text results in any tokens exceeding the specified length will be cut off and discarded. There are two options for truncation **Right** which is the default and **left** which can be passed as an argument into the tokenizer if required.
- **Left+Right Truncate the middle:** For any document that exceeds the maximum sequence length we take the first 128 tokens of the document and the last 382, these segments are then concatenated, taking 510 tokens in total, leaving room for BERT special tokens. Any text/tokens in the document that fall in between these values are removed.
- **Hierarchical text pre-processing – mean pooling:** hierarchical text pre-processing which involves splitting the text into 510 length segments. The model individually processes each of the document segments, and to get the classification results for a document in its entirety, we apply mean pooling across the multiple document segments.

UNIVERSITY OF LEEDS

Experimental Results

DOCUMENT CLASSIFICATION RESULTS

Model	Model Classification Results			
	Text Processing Strategy	Micro F1	Macro F1	ROC-AUC
BERT-base	Right Truncation	0.84	0.59	0.89
BERT-base	Left Truncation	0.82	0.52	0.88
BERT-base	Left+Right	**0.84**	**0.64**	**0.90**
BERT-base	Hierarchical Mean Pooling	0.84	0.61	0.89
Bio_ClinicalBERT	Right Truncation	0.82	0.52	0.88
Bio_ClinicalBERT	Left Truncation	**0.84**	**0.67**	**0.89**
Bio_ClinicalBERT	Left+Right	0.84	0.62	0.89
Bio_ClinicalBERT	Hierarchical Mean Pooling	0.84	0.63	0.89
BiomedBERT	Right Truncation	0.88	0.69	0.92
BiomedBERT	Left Truncation	**0.89**	**0.74**	**0.93**
BiomedBERT	Left+Right	0.86	0.67	0.90
BiomedBERT	Hierarchical Mean Pooling	**0.90**	**0.74**	**0.93**

MACRO-F1 SCORES FOR CLASSIFICATIONS BY TOKEN LENGTH DISTRIBUTION

Text Pre-processing Strategy + Token Length Distribution	Macro F1 Scores for Token Length Evaluation		
	BERT-base	Bio_CBERT	BioMBERT
Right >=1000	0.57	0.52	0.66
Right >=512 <1000	0.60	0.53	0.72
Right <512 >=250	0.60	0.52	0.70
Right <250	0.57	0.51	0.68
Left>=1000	0.51	0.60	0.72
Left >=512 <1000	0.53	0.67	**0.76**
Left <512 >=250	0.52	0.69	**0.77**
Left <250	0.51	0.66	0.72
Left+Right >=1000	0.58	0.60	0.62
Left+Right >=512 <1000	0.65	0.63	0.70
Left+Right <512 >=250	0.65	0.62	0.68
Left+Right >250	0.62	0.61	0.65

Conclusion

- We found performance increases using domain trained models over a generic model with a standard vocabulary.
- There are performance differences between domain trained models – not all model vocabularies are created equal.
- We observed that text pre-processing methods which use just the end of the pathology reports were most favourable to the clinical models – and some pathology reports contain a summary of key points at the end, which could explain this result.
- Documents far exceeding the maximum input length do suffer performance losses, but… also much shorter documents also suffer the same fate.

Future Work

- Applying multi-task learning to BERT models – in place of multi-label or single instance training.
- Only a subset of document label features were used in this study, there is potential for further analyses with a wider set of labels. Investigating a single cancer type with more breadth and depth e.g., looking at cancer specific bio markers and tumour sizes etc.
- BERT models are Deep Learning model architectures that are somewhat of a black box and investigating the models output using explainability methods is a next step for this research.

UNIVERSITY OF LEEDS

Thank you for listening!