IDENTIFY KEY PHRASES IN PATIENT NOTES USING NATURAL LANGUAGE PROCESSING (NLP)

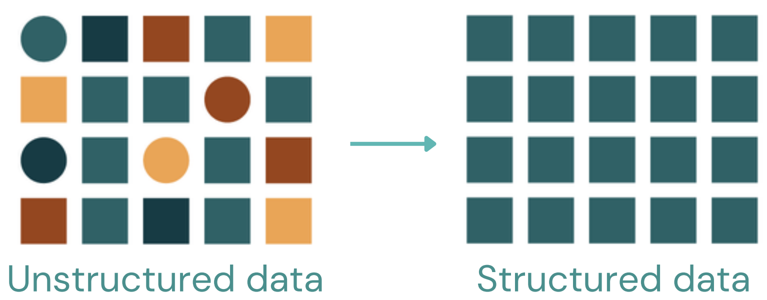
Cheng-Liang Lu, Rebecca Di Francesco, Irene Campillo Pereda
TECHNICAL UNIVERISTY OF DENMARK (DTU)

BACKGROUND.

NLP offers the ability to convert the unstructured data in personal health records into structured databases with standardised and international formats.

Advantages:

- accelerate research processes,
- help medical professionals make better decisions,
- build a clinical research network that enables collaboration between clinical research centres.



Entity recognition can be used to accurately identify information needed in unstructured personal health records. This would be the basis for developing a tool to assist professionals in decision making: for instance, a program that takes a patient record as input and outputs the diseases that the patient is most likely to have.

CHALLENGE.

The **NBME - Score Clinical Patient Notes Kaggle Competition** aims to develop a automated method for identifing clinical features in a patient note.

CLINICAL FEATURES	PATIENT NOTE

`45-year`
`Female`
`anxious-OR-nervous`
`No-depressed-mood`
`decreased-appetite`
`weight-stable`
`lack-of-thyroid-sympoms`
`insomnia`
	`45 yo F who has a +a few weeks of new onset nervousness. She states she had a decrease in appetite for the last week. She denies weight change, heat/cold intol., flushing, tremulousness, diarrhea.`

Figure 2 : Clinical features and their expressions within an example patient note.

DATA SET.

The data set consists of 43,985 clinical patient notes written by 35,156 examinees during the USMLE® Step 2 Clinical Skills examination. In this exam, examinees interact with standardized patients. For each encounter, an examinee writes a patient note, which is then scored by physician raters using a rubric of clinical features that should be present in the patient note.

MODEL ARCHITECTURE.

Since we did not have enough data to train a NLP model from scratch we did *transfer learning*.
Our model consists on a pre-trained **BERT model** + FC layer.

But, why **BERT**?

1. **Built-in knowledge about language.**
   - It is less computationally expensive to fine-tune a pre-trained BERT model than to train a model from scratch on our specific task, that would require lot of data !
   - BERT-base: trained on BooksCorpus, ~800 million words!
2. **Ability understand context thanks to bidirectional training.**
   - BERT training approaches:Masked Language Modelling (MLM) & Next sentence prediction (NSP)

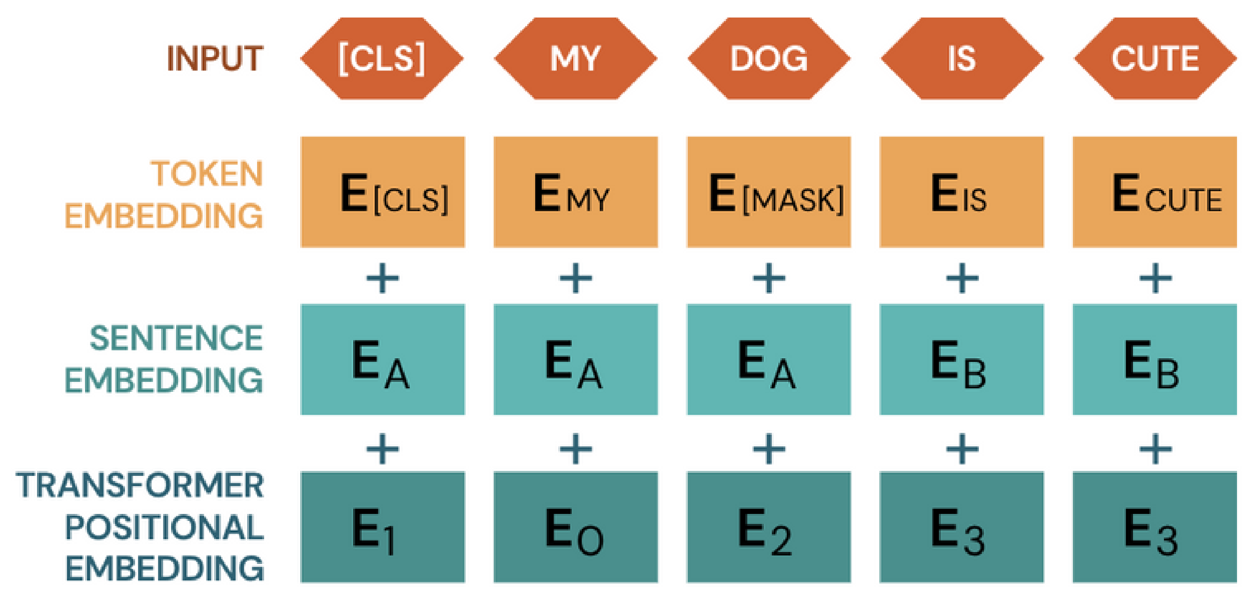


Figure 1 : BERT input representation.

RESULTS.

We followed BERT paper guidelines in hyper-parameter tuning in order to find the best model for our task.

We conducted experiments considering the following recommended hyperparameters:

- **batch size**: 16, 32.
- **number of epoches**: 2, 3, 4.
- **learning rate (Adam)**: 5e-5, 3e-5, 2e-5.

We refer to table below for examples of the top models.

batch size	number of epoches	learning rate	acc	loss	F1
16	4	5e-5	0.958	0.187	0.768
16	4	3e-5	0.950	0.233	0.694
16	4	2e-5	0.940	0.307	0.639

Table 1 : Dev Results of 3 of the top models.

REFERENCES.

- Devlin, J. et al. (2018) Pre-training of Deep Bidirectional Transformers for Language Understanding, Google AI language.
- Horev, R. (2018) Bert explained: State of the art language model for NLP, Towards Data Science.
- Pogiatzis, A. (2019) NLP: Contextualized word embeddings from bert, Towards Data Science.
- Vaswani, A. et al. (2017) Attention is all you need, Google Brain.
- Wei, J. (2020) Bert: Why it's been revolutionizing NLP, Towards Data Science.
- Yaneva, V. et al. (2022) The USMLE® step 2 clinical skills patient note corpus.
- Xi Yang et al. (2022) GatorTron: A Large Language Model for Clinical Natural Language Processing