IntelliCheck: Diagnosis Prediction Model

April 17, 2023

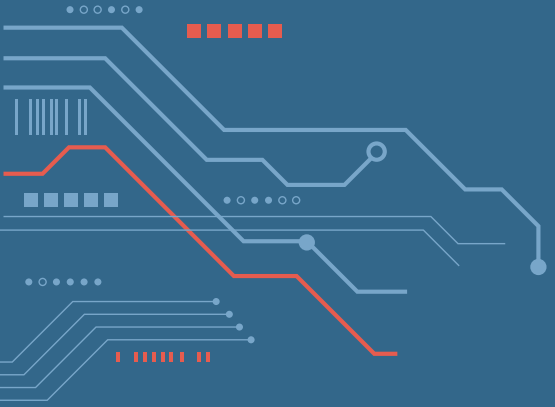
TABLE OF CONTENTS

01 Objectives

02 Methodology

03 Results & Analysis

04 Conclusions



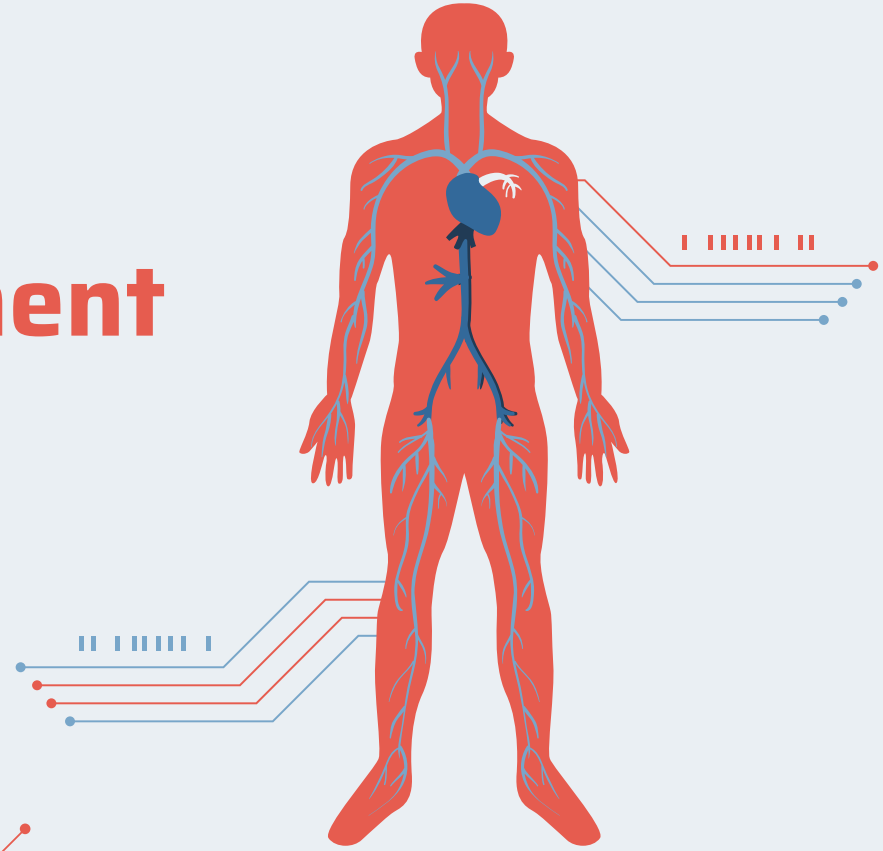
01

Objectives

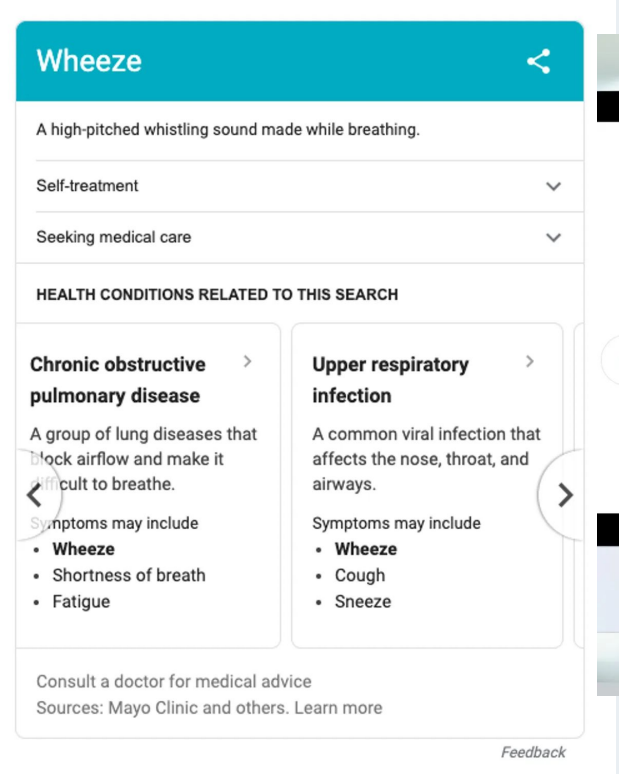
Problem Statement

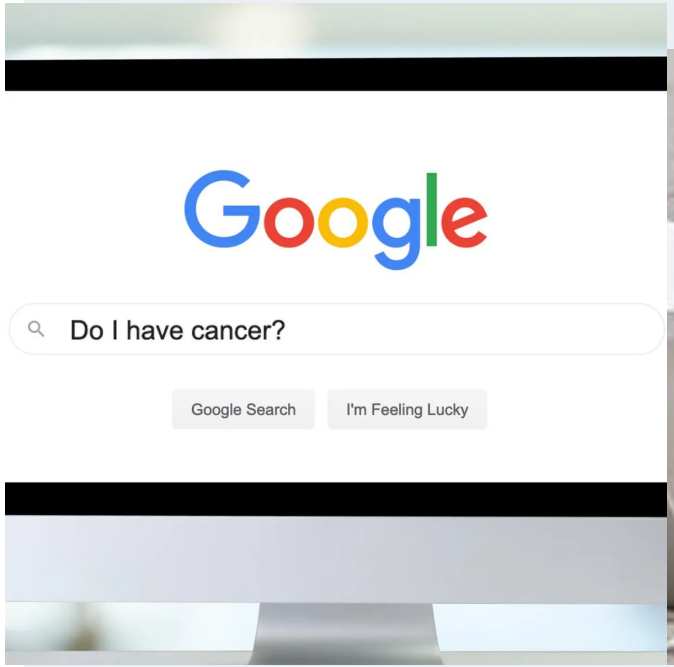
Doctors: Spend lot of time on diagnosis, leaving less time for treatment/patient care

Patients: Have to schedule doctor visit (and spend money) just to get assessment of illness severity before treatment



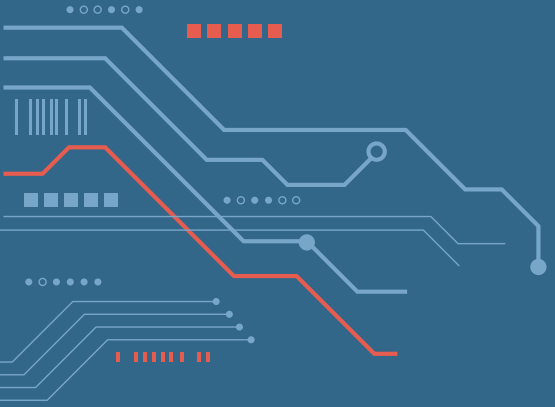
Existing Solutions





Literature Review Summary

- BioBERT
  - Trained on biomedical corpus such as PubMed and PMC articles
  - Tasks: named entity recognition, relation extraction, and question answering
- Med-BERT
  - Trained on electronic health records
  - Tasks: disease prediction
    - The prediction of heart failure among patients with diabetes
    - The prediction of onset of pancreatic cancer
- CORe
  - Trained on patient discharge records as well as biomedical articles (PMC, Wikipedia, etc)
  - Tasks: diagnosis prediction, procedure prediction, in-hospital mortality prediction, and length of stay prediction



02

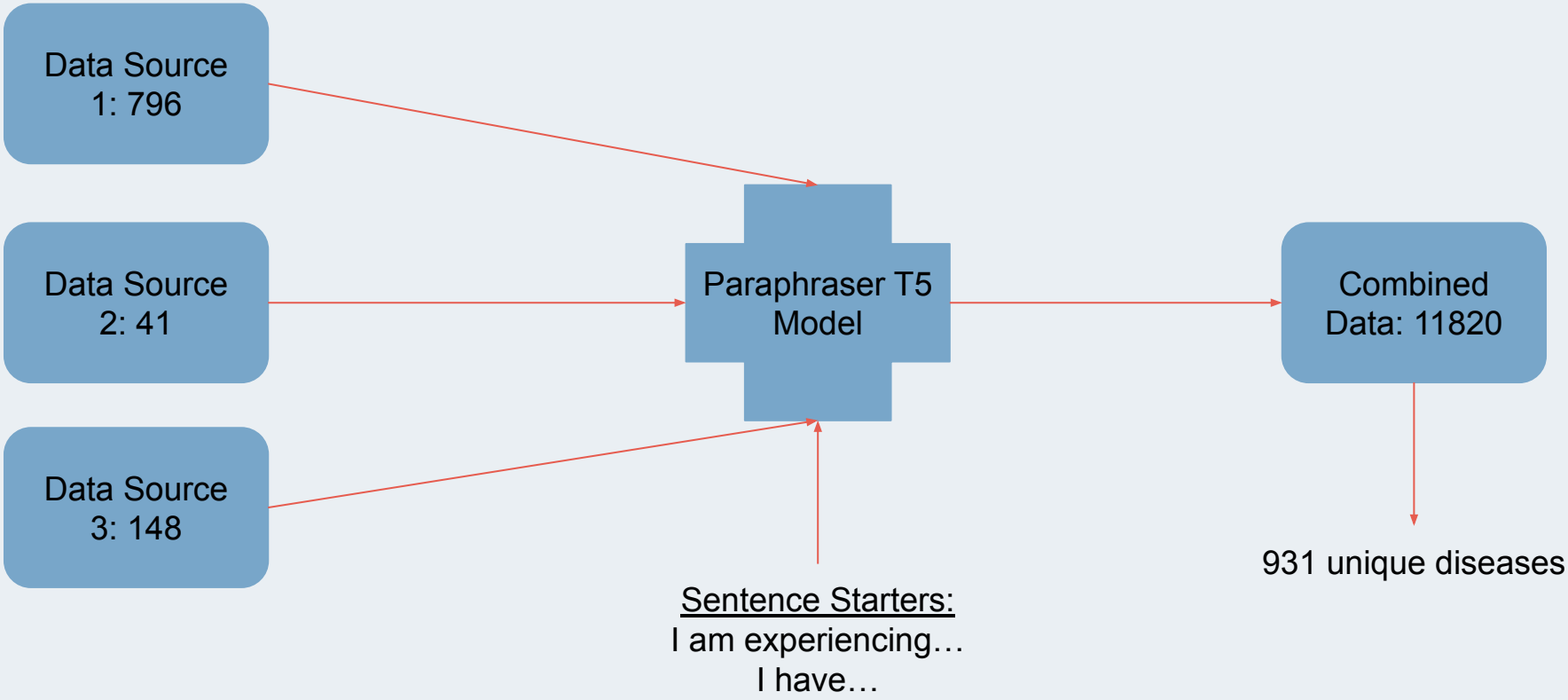
Methodology

Data Collection

3 Main Datasets:

- Kaggle: Diseases and their Symptoms
- Kaggle: Disease Symptom Prediction
- Columbia University: Disease-Symptom Knowledge Database

Data Paraphrasing Pipeline



T5 Model: Vamsi/T5_Paraphrase_Paws

Raw Data

	Unnamed: 0	Disease	Symptoms
0	0	fungal infection	itching, skin rash, nodal skin eruptions, d...
1	1	allergy	continuous sneezing, shivering, chills, wa...
2	2	gerd	stomach pain, acidity, ulcers on tongue, v...
3	3	chronic cholestasis	itching, vomiting, yellowish skin, nausea, ...
4	4	drug reaction	itching, skin rash, stomach pain, burning m...
5	5	peptic ulcer diseae	vomiting, loss of appetite, abdominal pain,...
6	6	aids	muscle wasting, patches in throat, high fev...
7	7	diabetes	fatigue, weight loss, restlessness, lethar...
8	8	gastroenteritis	vomiting, sunken eyes, dehydration, diarrhoea
9	9	bronchial asthma	fatigue, cough, high fever, breathlessness...
10	10	hypertension	headache, chest pain, dizziness, loss of b...
11	11	migraine	acidity, indigestion, headache, blurred an...
12	12	cervical spondylosis	back pain, weakness in limbs, neck pain, d...
13	13	paralysis	vomiting, headache, weakness of one body si...
14	14	jaundice	itching, vomiting, fatigue, weight loss, h...

Paraphrased Data

	Unnamed: 0	Disease	Symptoms
0	0	fungal infection	I have itching, skin rash, nodal skin erupti...
1	1	allergy	I have continuous sneezing, shivering, chil...
2	2	gerd	I have stomach pain, acidity, ulcers on ton...
3	3	chronic cholestasis	I have itching, vomiting, yellowish skin, n...
4	4	drug reaction	I have itching, skin rash, stomach pain, bu...
5	5	peptic ulcer diseae	I have vomiting, loss of appetite, abdomina...
6	6	aids	I have muscle wasting, patches in throat, h...
7	7	diabetes	I have fatigue, weight loss, restlessness, ...
8	8	gastroenteritis	I have vomiting, sunken eyes, dehydration, ...
9	9	bronchial asthma	I have fatigue, cough, high fever, breathl...
10	10	hypertension	I have headache, chest pain, dizziness, lo...
11	11	migraine	I have acidity, indigestion, headache, blu...
12	12	cervical spondylosis	I have back pain, weakness in limbs, neck p...
13	13	paralysis	I have vomiting, headache, weakness of one ...
14	14	jaundice	I have itching, vomiting, fatigue, weight l...

Training Overview

Train/Val/Test Split

Used stratified
75-15-10 split

Prelim Model Training

Trained 5 models for
5 epochs to gauge
performance

(LR: $2x10^{-5}$, Opt: AdamW, LR Decay:
Linear w/ Rate of 0.01, Batch Size: 16)

Final Model Training

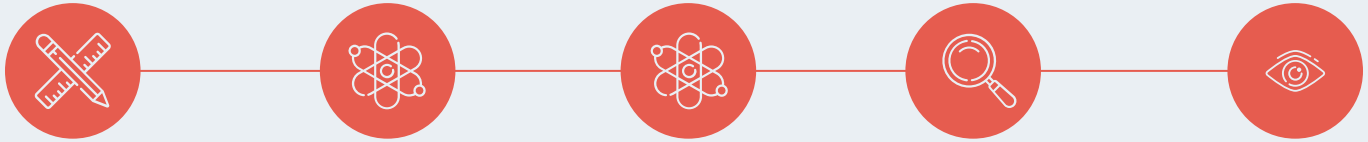
Trained best model
for 5 more epochs

Final Model Testing

Evaluated model on
withheld testing data

Inference (Demo)

Applied model to
pipeline



Prelim Transformer Models

Bio-ClinicalBERT

Trained on notes from EHRs for natural language inference and named entity recognition tasks

Monlogg/Biobert

Trained on PubMed articles for masked language modeling

DMIS-Lab/Biobert

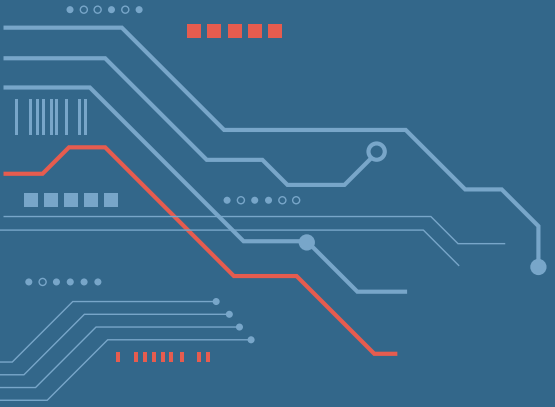
Fine tuning of BioBERT by Data Mining and Information Systems Lab at Korea University

Bioinformer-8L

Trained on a biomedical corpus of PubMed and PMC articles for masked language modeling

Microsoft/BiomedNLP

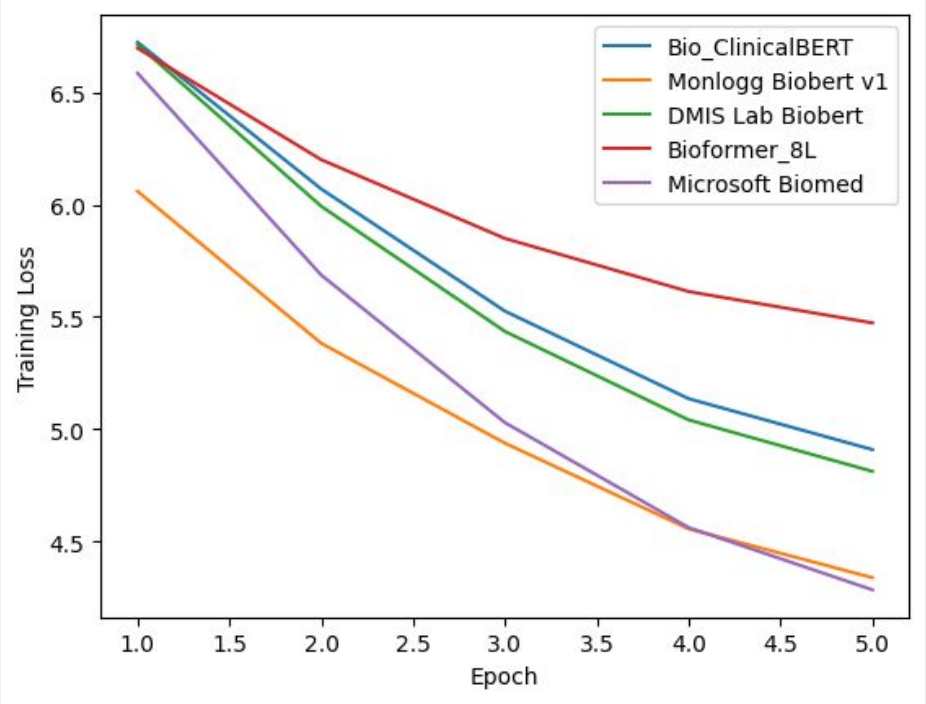
Trained PubMed abstracts and PMC articles for many NLP tasks including understanding and reasoning



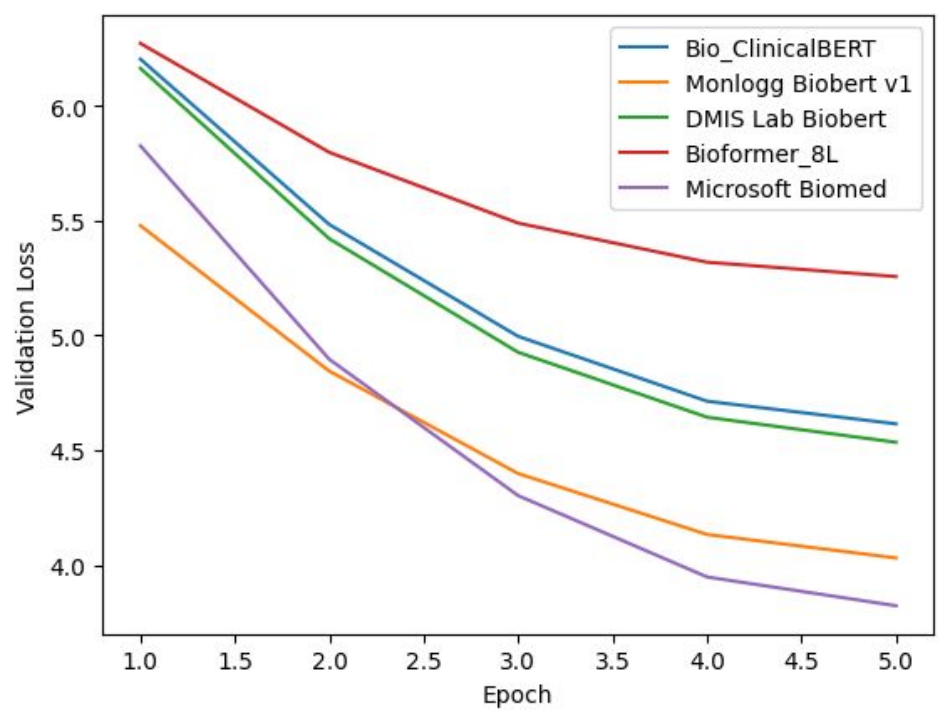
03

Results & Analysis

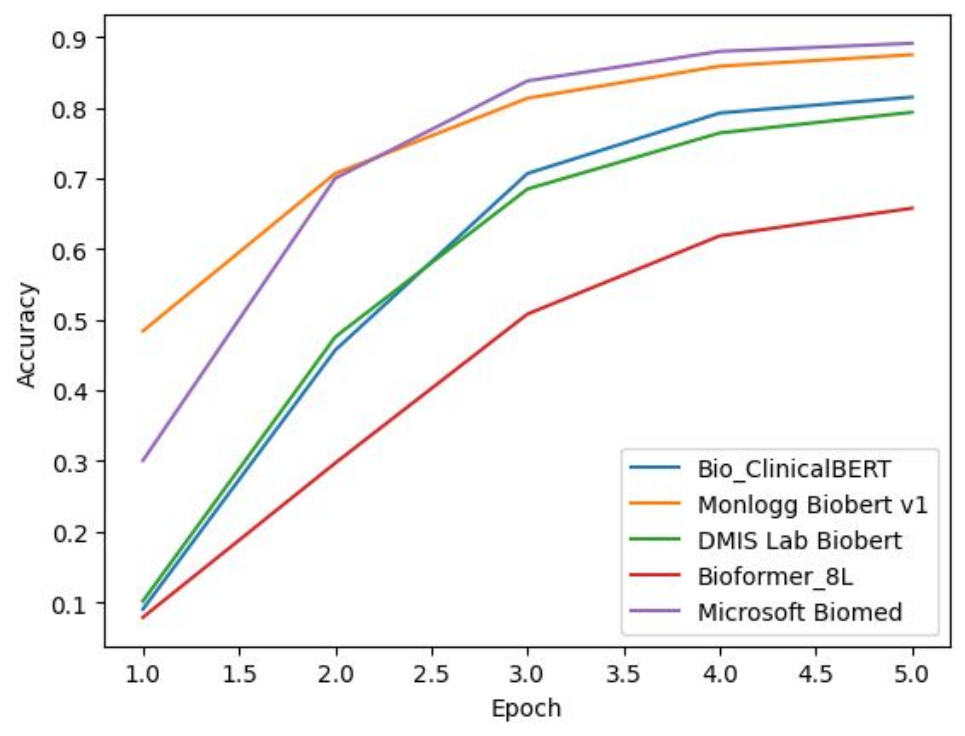
Training Loss Results



Validation Loss Results



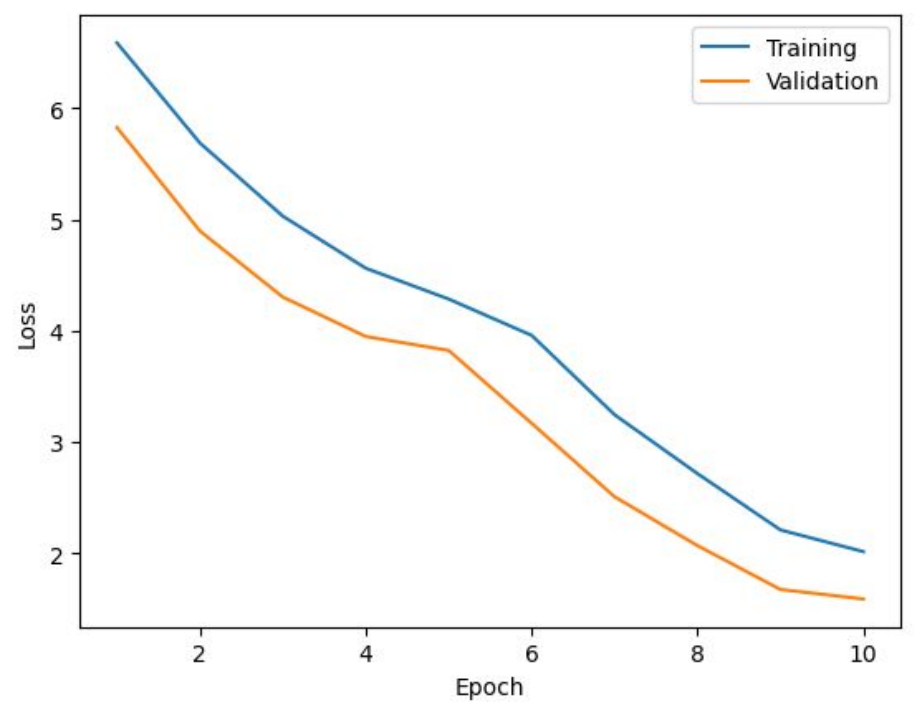
Validation Accuracy Results



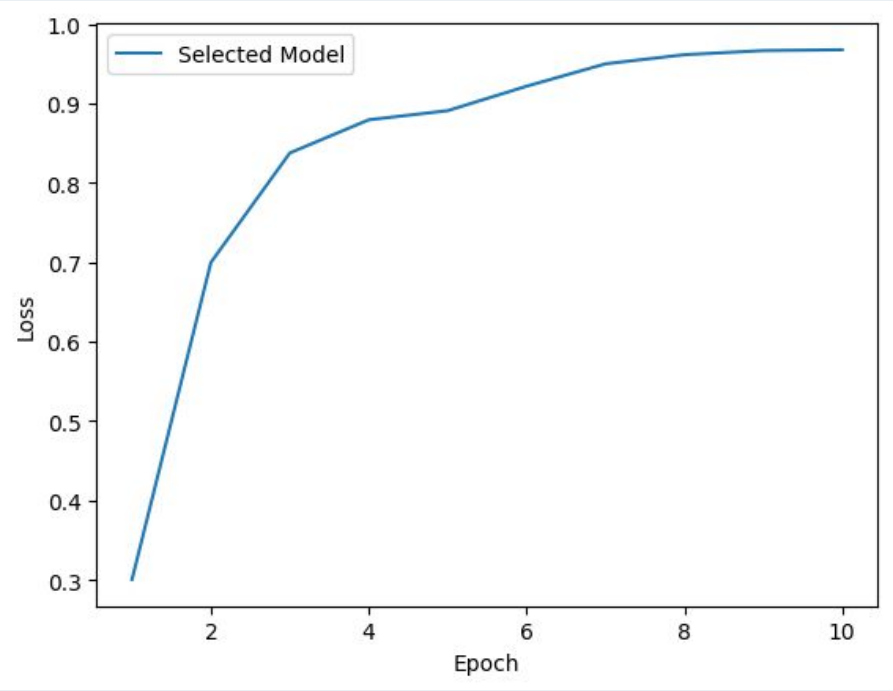
Analysis of Results

- Microsoft/BiomedNLP and Monlogg/BioBERT are top performers on 5 epochs
- After 5 epochs, however, Microsoft outperformed Monlogg
- Monlogg seems to plateau/converge quicker than Microsoft
- Thus, chosen model is **Microsoft BiomedNLP**

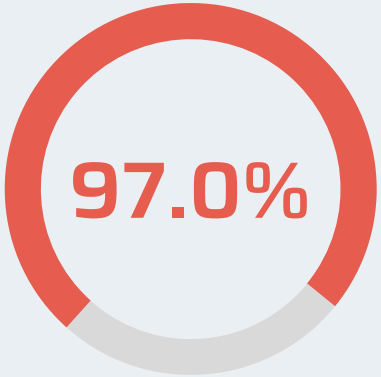
Microsoft-BiomedNLP Loss



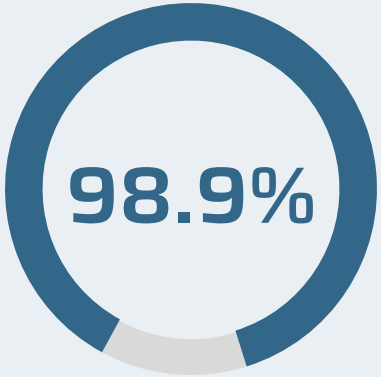
Microsoft-BiomedNLP Validation Accuracy



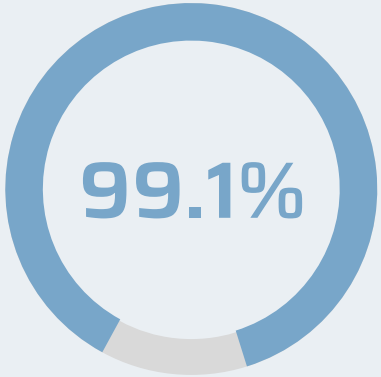
Microsoft-BiomedNLP Testing Results



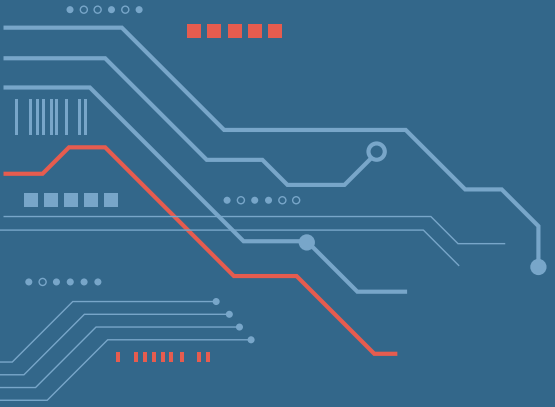
Test Accuracy



Test Top 3 Accuracy



Test Top 5 Accuracy



04

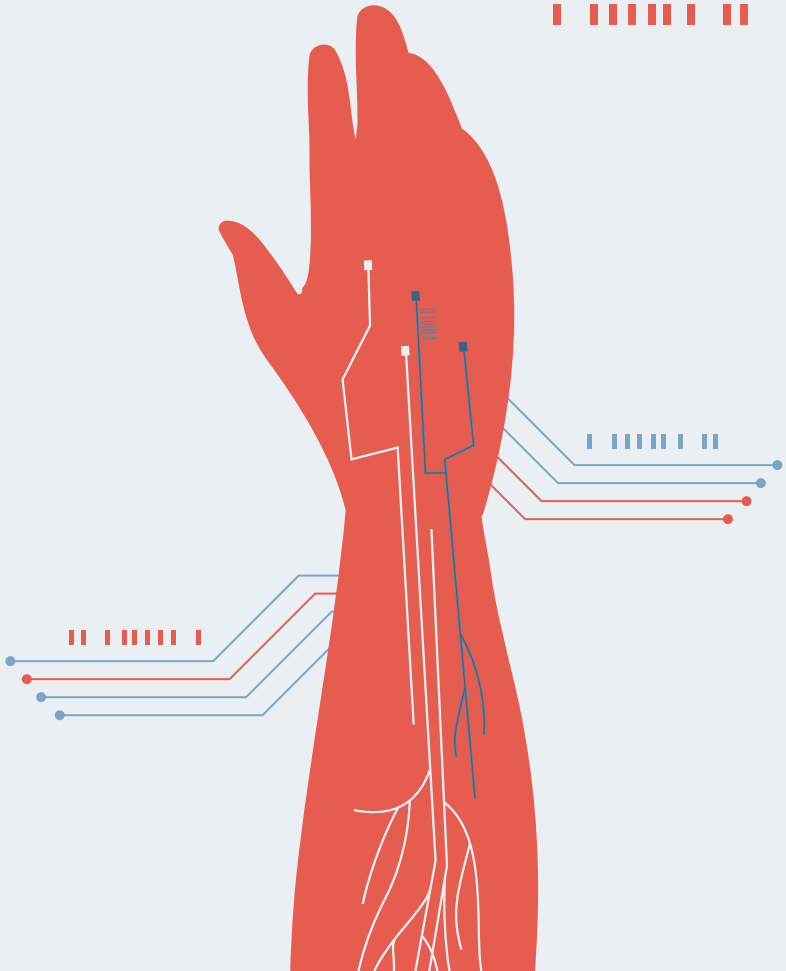
Conclusions

Limitations & Future Work

- Scope and availability of training data (931 labels/conditions)
  - Difficult to find adequate symptoms and disease data
  - Need to get creative with collecting/generating more data
- Lack of hyperparameter tuning
  - Use grid search or similar methodology to optimize hyperparameters
- Performance on zero-shot/few-shot classification?
  - Still need to see how well model can perform on out-of-scope data
- Disclaimer: Model should not replace the advice of licensed professionals, especially in severe circumstances
  - Intended to aid individuals with minor illnesses or determine what next steps are needed

CONCLUSIONS

Despite the previously mentioned limitations, the results are promising. With incorporation of more data, our fine-tuned model has the potential to alleviate extra stress (due to mild or non-emergency requests) on the healthcare system by providing accurate diagnosis predictions.

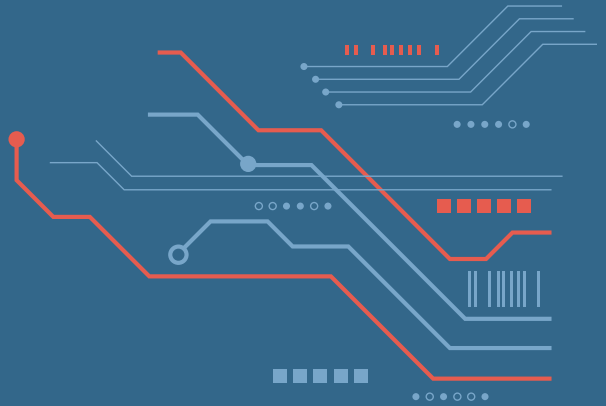


Works Cited

- https://huggingface.co/course/chapter2/6?fw=pt
- https://huggingface.co/course/chapter3/4?fw=pt
- https://huggingface.co/docs/transformers/v4.27.2/en/tasks/sequence_classification#preprocess
- https://medium.com/@eijaz/introduction-to-text-classification-using-transformers-be55d49cdd88
- https://towardsdatascience.com/transformer-models-for-custom-text-classification-through-fine-tuning-3b065cc08da1
- https://levelup.gitconnected.com/how-to-fine-tune-an-nlp-classification-model-with-transformers-and-huggingface-1a2c0ea79c2
- https://adaa.org/learn-from-us/from-the-experts/blog-posts/consumer/health-anxiety-what-it-and-how-beat-it
- https://www.cnet.com/health/medical/google-medical-symptoms/
- https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext
- https://huggingface.co/bioformers/bioformer-8L
- https://huggingface.co/dmis-lab/biobert-v1.1
- https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
- https://huggingface.co/monologg/biobert_v1.1_pubmed

THANKS

Do you have any questions?



Please keep this slide for attribution.

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.