Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Using Natural Language Processing (NLP) to predict colon cancer from discharge summaries

CEB workshop

25 November 2022 13:00 – 15:00

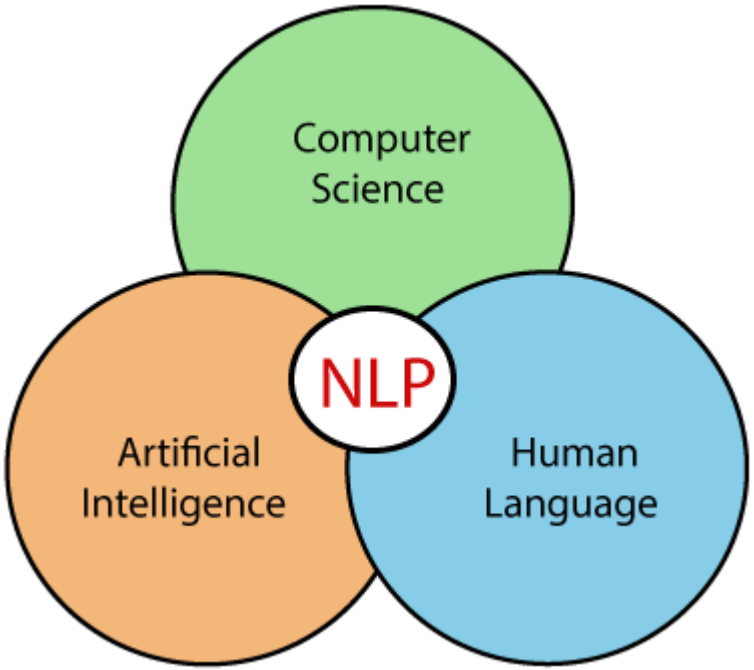
Wanchana Ponthongmak

Presentation available here!
shorturl.at/eV348

What is Natural Language Processing (NLP)?

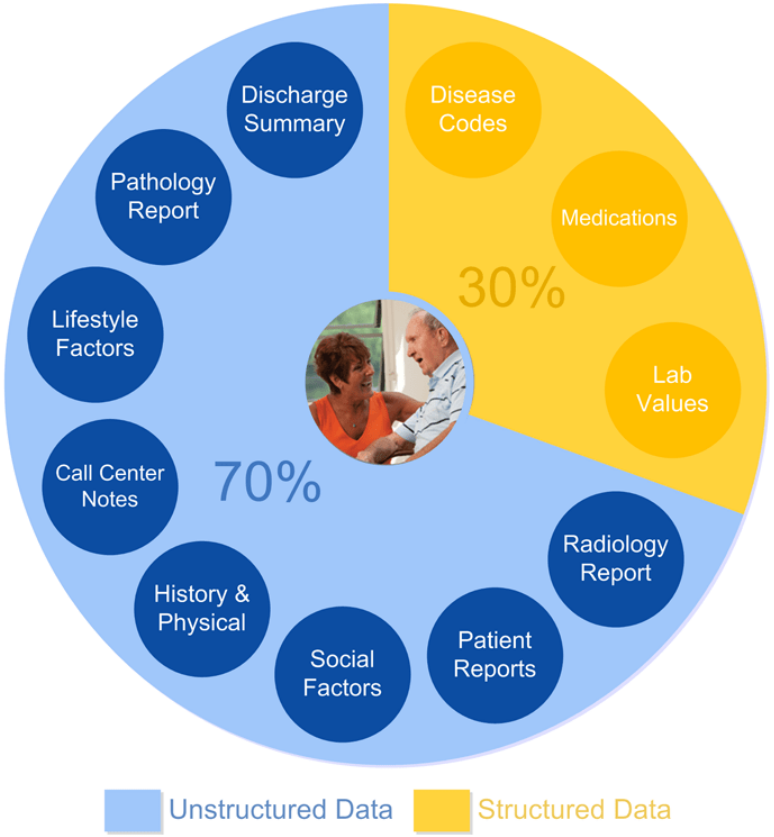
- NLP is a multi-disciplinary fields, which is a part of computer science, human language, and artificial intelligence (AI).

- It offers computers ability to analyze, understand, and utilize human languages as they are spoken and written,
  - Textual form
  - Vocal form



Why NLP?

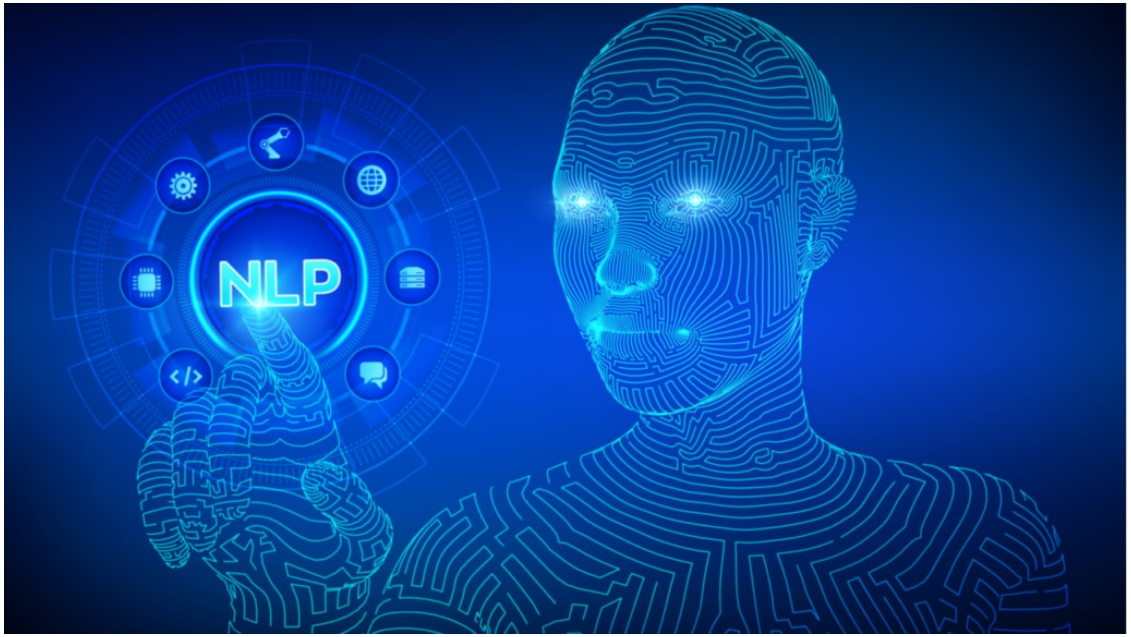
- Text data are extremely rich source of information
- But extracting insights from them can be hard and time-consuming, due to its unstructured nature.
- ~70% of data in hospitals are unstructured data



https://www.altexsoft.com/blog/nlp-healthcare/

Why NLP?

- But, thanks to advances in NLP, machine learning (ML), deep learning (DL) and AI, sorting text data is getting easier.

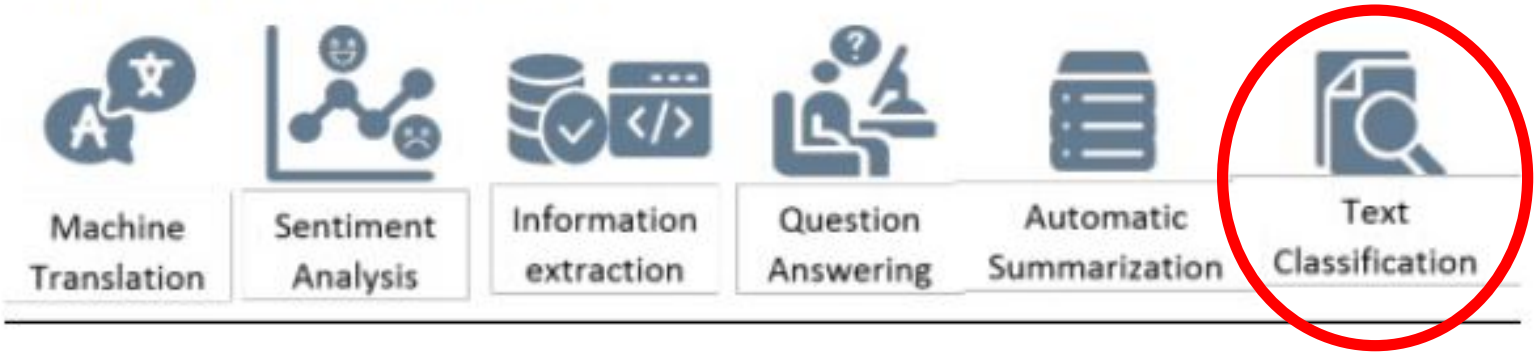


https://online.york.ac.uk/the-role-of-natural-language-processing-in-ai/

Why NLP?

NLP Applications

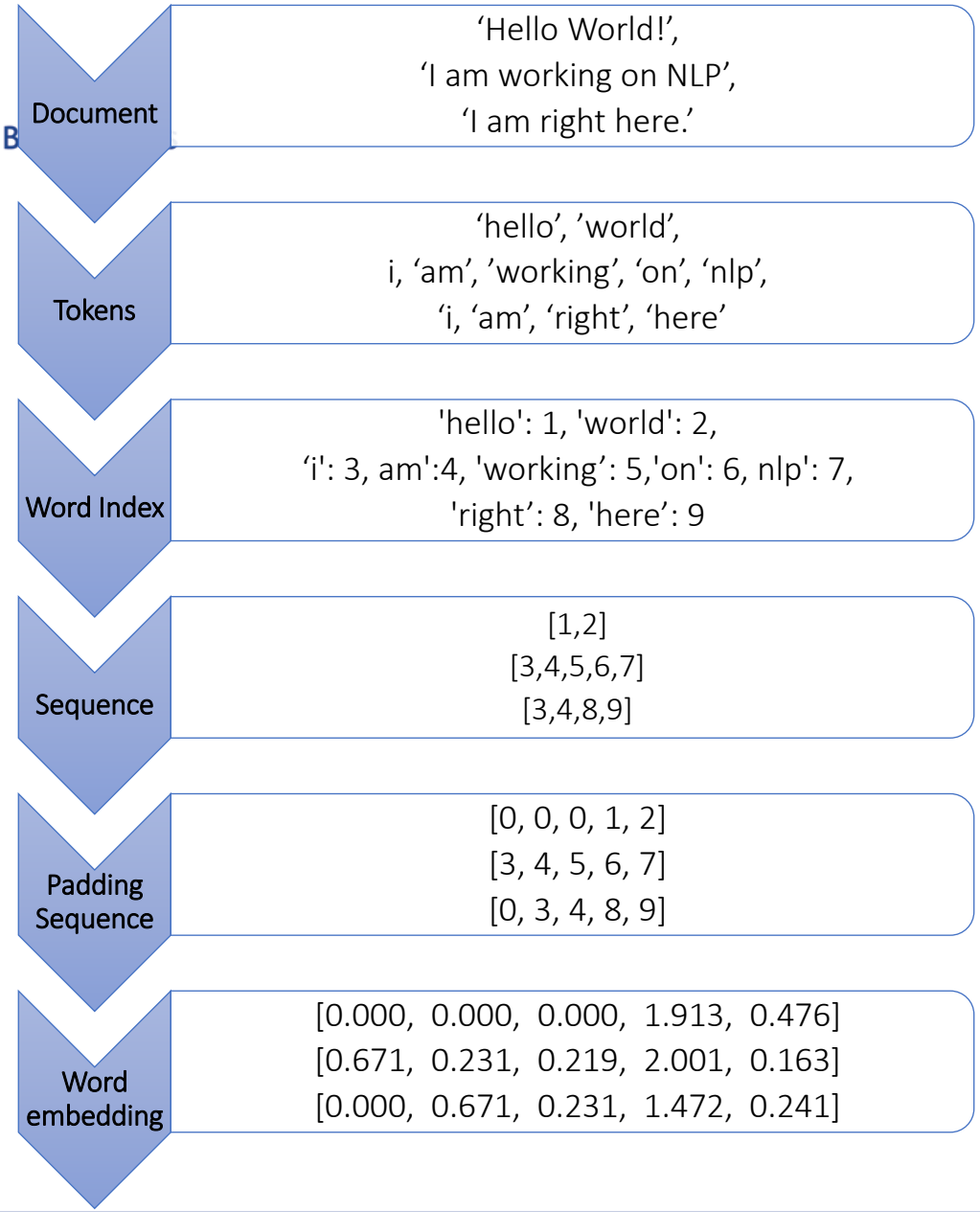
Frequent applications of NLP are



https://www.forbesindia.com/article/weschool/natural-language-processing-a-breakthrough-technology-in-healthcare/67661/1

How to deal with text data?

- Structurize text data
  - vectorize to numerical vector with the same size

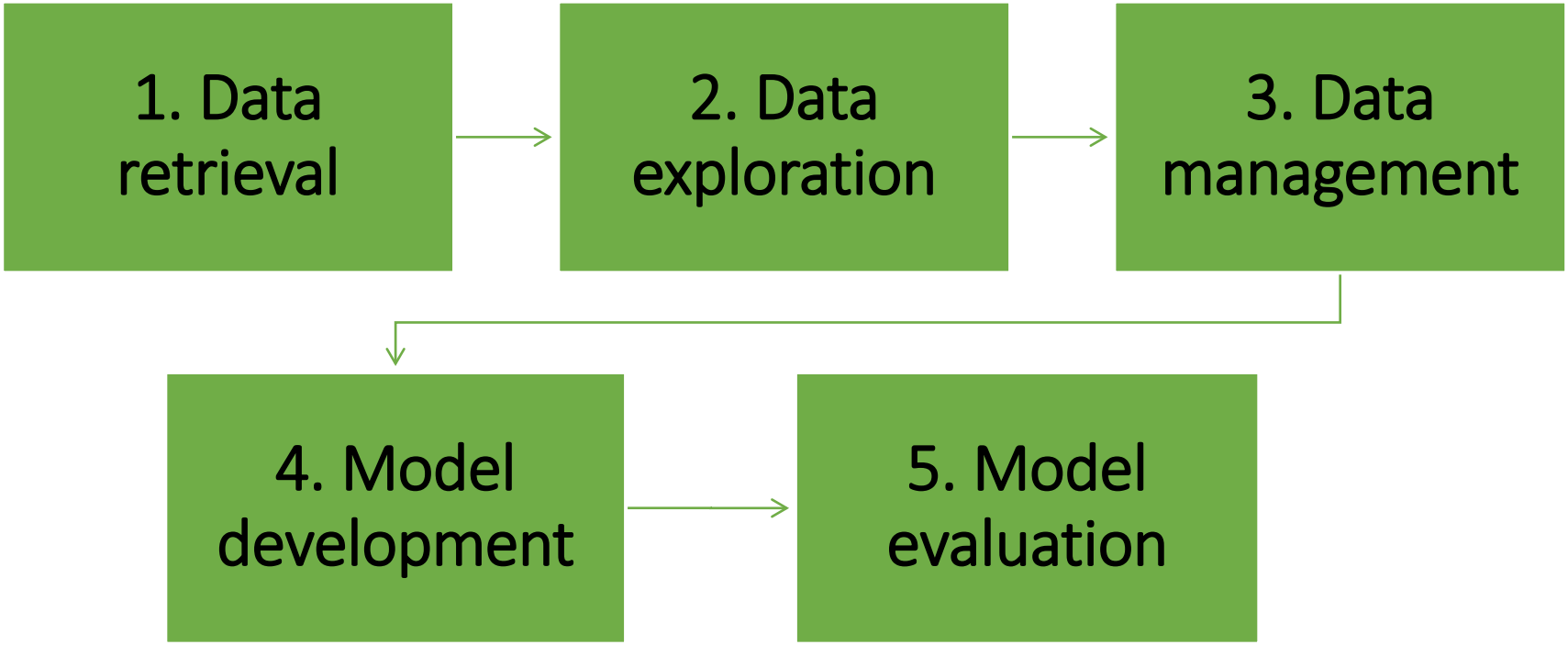


Objectives of the workshop

To develop models to predict colon cancer from discharge summary using NLP with machine learning

To evaluate the performance of developed models

Framework


1. Data retrieval
2. Data exploration
3. Data management
4. Model development
5. Model evaluation

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Data Retrieval

Study design and setting

A cross-sectional study from
Department of Medicine, Ramathibodi Hospital
1st January 2015 – 31st December 2019,

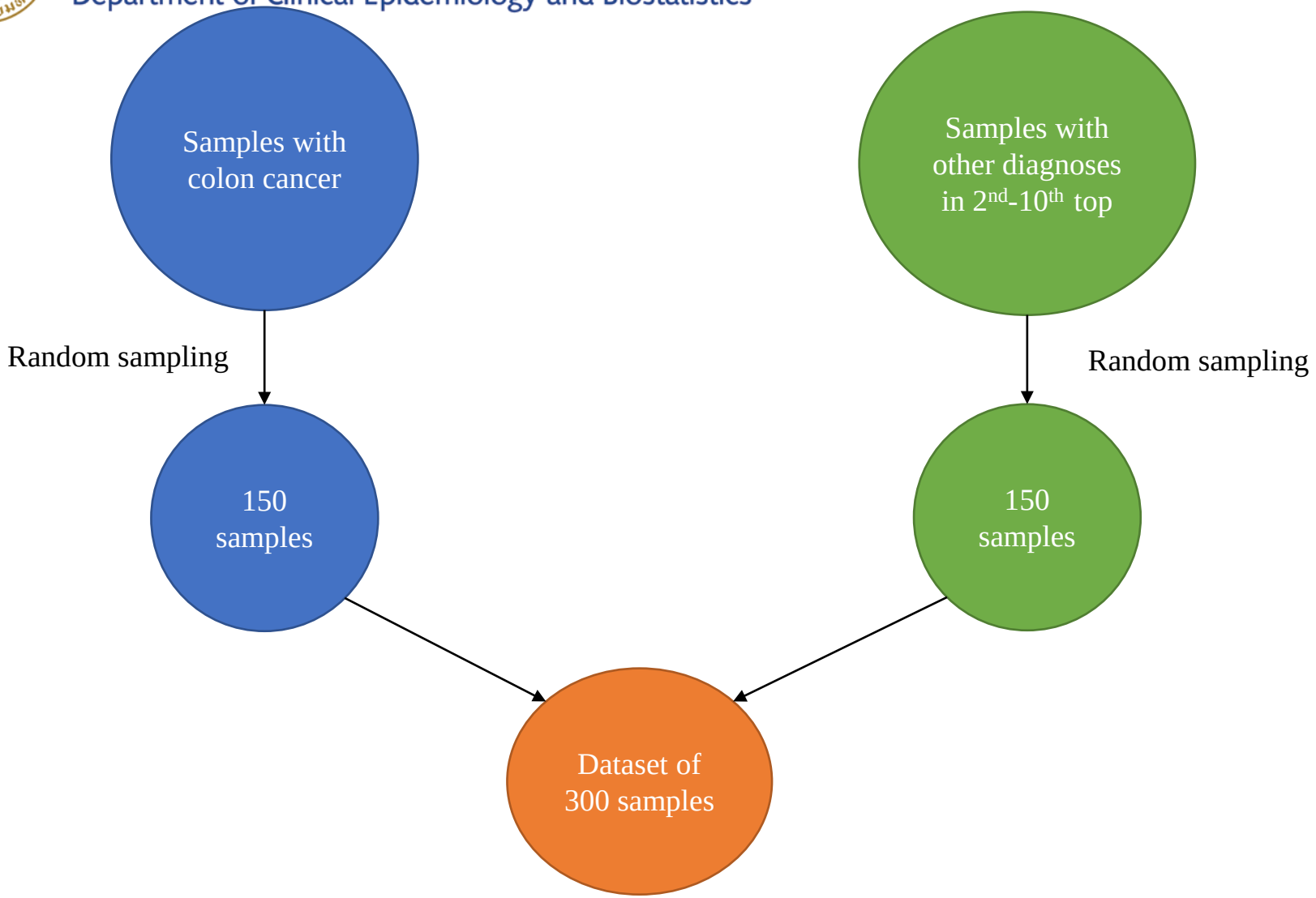
Ramathibodi Human Research Ethics Committee
(COA.MURA2020/1152).

Inclusions

- Audited charts
- ≥18 years old
- Top 10th diseases

Exclusions

- Missing diagnosis codes
- Missing clinical notes


Samples with colon cancer
Samples with other diagnoses in 2nd-10th top
Random sampling
Random sampling
150 samples
150 samples
Dataset of 300 samples

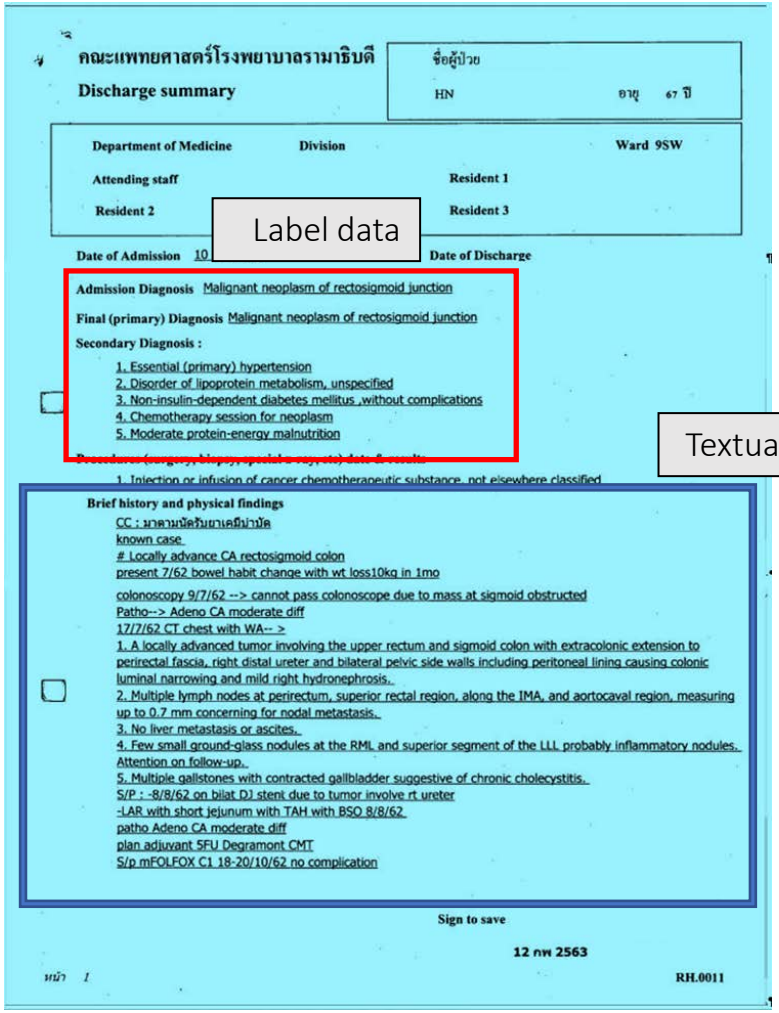
Data Exploration

The data exploration applied descriptive statistics and data visualization techniques to understand the characteristic of the study datasets

- Features exploration

- Data characteristics
  - # of patients
  - # of documents
  - Age / gender
  - # of colon CA
  - Vocabulary size
  - # of tokens per documents

hn	an	age	gender	pdx	brief_raw	course_raw	label
wpLClmZowqrCo	wpLCmGpmwq3C	49	2	C20	known case CA rectum Dx. ไ	CA rectosigmoid colon c ova	1
wpPCmmNqwqv	wpLCmWRswq3C	67	1	C20	case CA lower rectum S/P A	#recurrent CA lower rectum	1
wpXCk2lpwrHCp	wpLCmWtswrHCı	45	2	C20	known case ca rectum c live	# CA rectal with liver metast	1
wpbCl2pswq7Cp	wpLCm2RswqzCr	81	2	C20	Known case CA Rectum with	case CA Rectum with liver m	1
wpbCl2puwqzCn	wpLCm2Vuwq_C	62	2	C20	Known case : CA rectum (T3	S/P mFOLFOX6 1st cycle 3-4,	1
wpbCk2huwrLCo	wpLCmWhswqrC	67	2	C20	Known case : CA upper rectı	1. Recurrent CA rectum เป็น	1
wpbCl2NnwqvCn	wpLCmmxlwrLCc	71	1	C20	known case CA rectum S/P l	#CA rectum S/P LAR start mF	1
wpPCl2xuwrLCoM	wpLCmmlswrDCp	57	1	C20	#Known case CA rectum pre	#Known case CA rectum T3M	1
wpbCk2duwq7Cr	wpLCmWhmwrLC	65	2	C20	Known case CA lower rectur	# CA lower rectum with LN r	1
wpLCkmdswqnCc	wpLCmGlnwrHCc	68	2	C20	Known case # CA rectum wi	# CA rectum with liver meta	1
wpXCm2luwq_Cr	wpLCmGtnwrHCr	56	1	C20	#U/D DM, HT, old CVA + yrs i	#U/D DM, HT, old CVA 10+ yı	1
wpbClmlqwq7Co	wpLCmmtrwqvCc	70	1	C20	U/D IHD S/P CAG 23/11/61 k	#CA rectosigmoid pT3N1M0	1
wpXCk2ptwqnCo	wpLCmmtlwq7Cr	76	1	C20	Known case # DLP # small bc	# DLP # small bowel GIST s/	1
wpbCl2Rmwq7Cp	wpLCmmtrwqzCc	57	1	C20	Known case CA upper rectur	#CA upper rectum with mul	1
wpbClWpowq7Cı	wpLCmmpuwrDC	46	1	C20	known case CA lower rectur	#CA lower rectum stage IIIb	1



คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี

Discharge summary

ชื่อผู้ป่วย
HN อายุ 67 ปี

Department of Medicine Division Ward 9SW
Attending staff Resident 1
Resident 2 Resident 3

Label data

Date of Admission 10 Date of Discharge

Admission Diagnosis Malignant neoplasm of rectosigmoid junction

Final (primary) Diagnosis Malignant neoplasm of rectosigmoid junction

Secondary Diagnosis :

1. Essential (primary) hypertension
2. Disorder of lipoprotein metabolism, unspecified
3. Non-insulin-dependent diabetes mellitus ,without complications
4. Chemotherapy session for neoplasm
5. Moderate protein-energy malnutrition

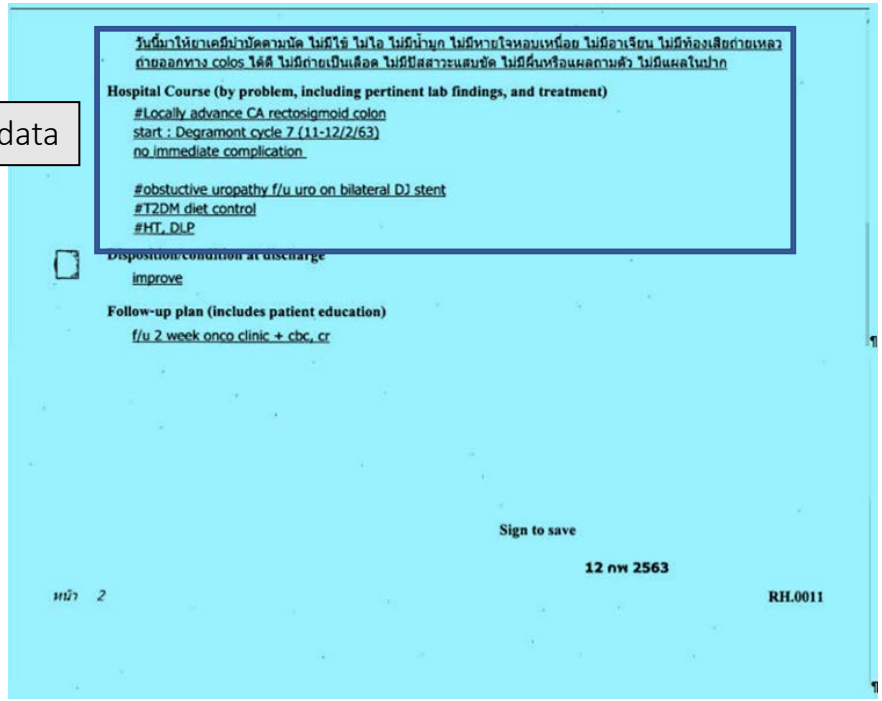
1. Injection or infusion of cancer chemotherapeutic substance, not elsewhere classified

Brief history and physical findings

CC : มาตามนัดรับยาเคมีบำบัด
known case
Locally advance CA rectosigmoid colon
present 7/62 bowel habit change with wt loss10kg in 1mo
colonoscopy 9/7/62 --> cannot pass colonoscope due to mass at sigmoid obstructed
Patho--> Adeno CA moderate diff
17/7/62 CT chest with WA-- >
1. A locally advanced tumor involving the upper rectum and sigmoid colon with extracolonic extension to perirectal fascia, right distal ureter and bilateral pelvic side walls including peritoneal lining causing colonic luminal narrowing and mild right hydronephrosis.
2. Multiple lymph nodes at perirectum, superior rectal region, along the IMA, and aortocaval region, measuring up to 0.7 mm concerning for nodal metastasis.
3. No liver metastasis or ascites.
4. Few small ground-glass nodules at the RML and superior segment of the LLL probably inflammatory nodules. Attention on follow-up.
5. Multiple gallstones with contracted gallbladder suggestive of chronic cholecystitis.
S/P : -8/8/62 on bilat DJ stent due to tumor involve rt ureter
-LAR with short jejunum with TAH with BSO 8/8/62
patho Adeno CA moderate diff
plan adjuvant 5FU Degramont CMT
S/p mFOLFOX C1 18-20/10/62 no complication

Sign to save
12 กพ 2563
หน้า 1 RH.0011

Textual data



วันนี้มาให้ยาเคมีบำบัดตามนัด ไม่มีไข้ ไม่ไอ ไม่มีน้ำมูก ไม่มีหายใจหอบเหนื่อย ไม่มีอาเจียน ไม่มีท้องเสียถ่ายเหลว ถ่ายออกทาง colos ได้ดี ไม่มีถ่ายเป็นเลือด ไม่มีปัสสาวะแสบขัด ไม่มีผื่นหรือผลตามตัว ไม่มีผลในปาก

Hospital Course (by problem, including pertinent lab findings, and treatment)

#Locally advance CA rectosigmoid colon
start : Degramont cycle 7 (11-12/2/63)
no immediate complication

#obstuctive uropathy f/u uro on bilateral DJ stent
#T2DM diet control
#HT, DLP

Disposition/condition at discharge

improve

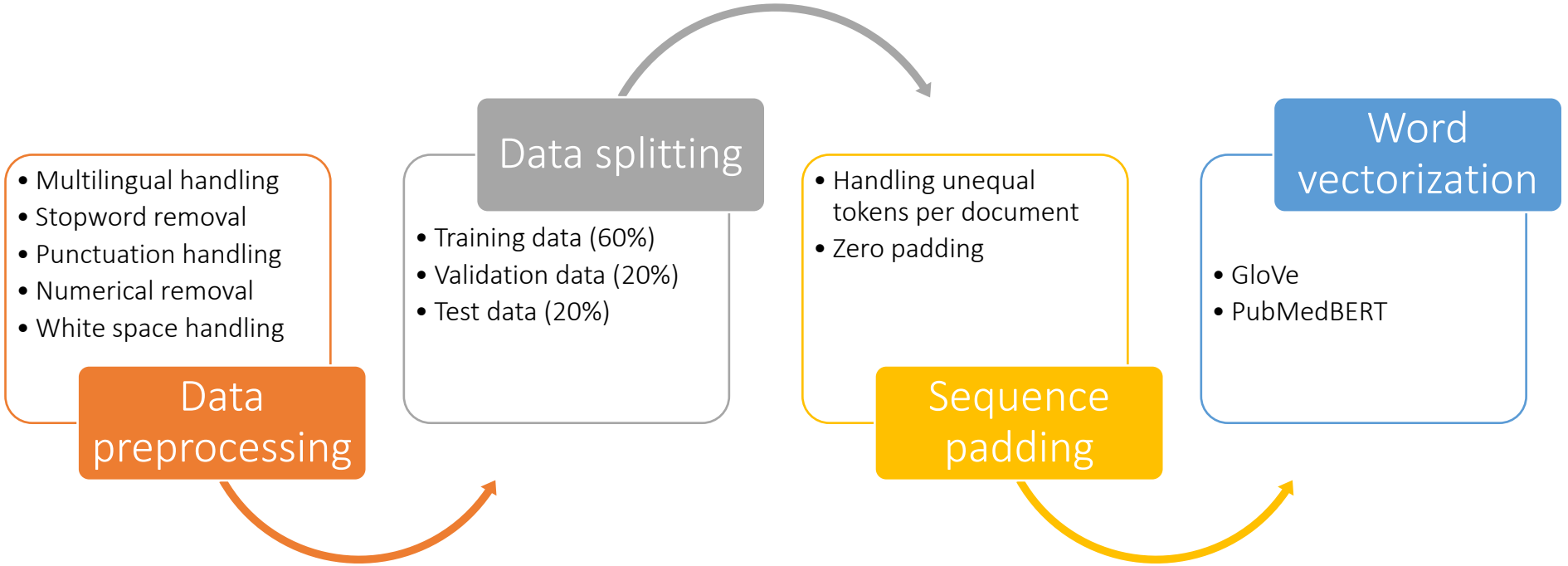
Follow-up plan (includes patient education)

f/u 2 week onco clinic + cbc, cr

Sign to save
12 กพ 2563
หน้า 2 RH.0011

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Data Management

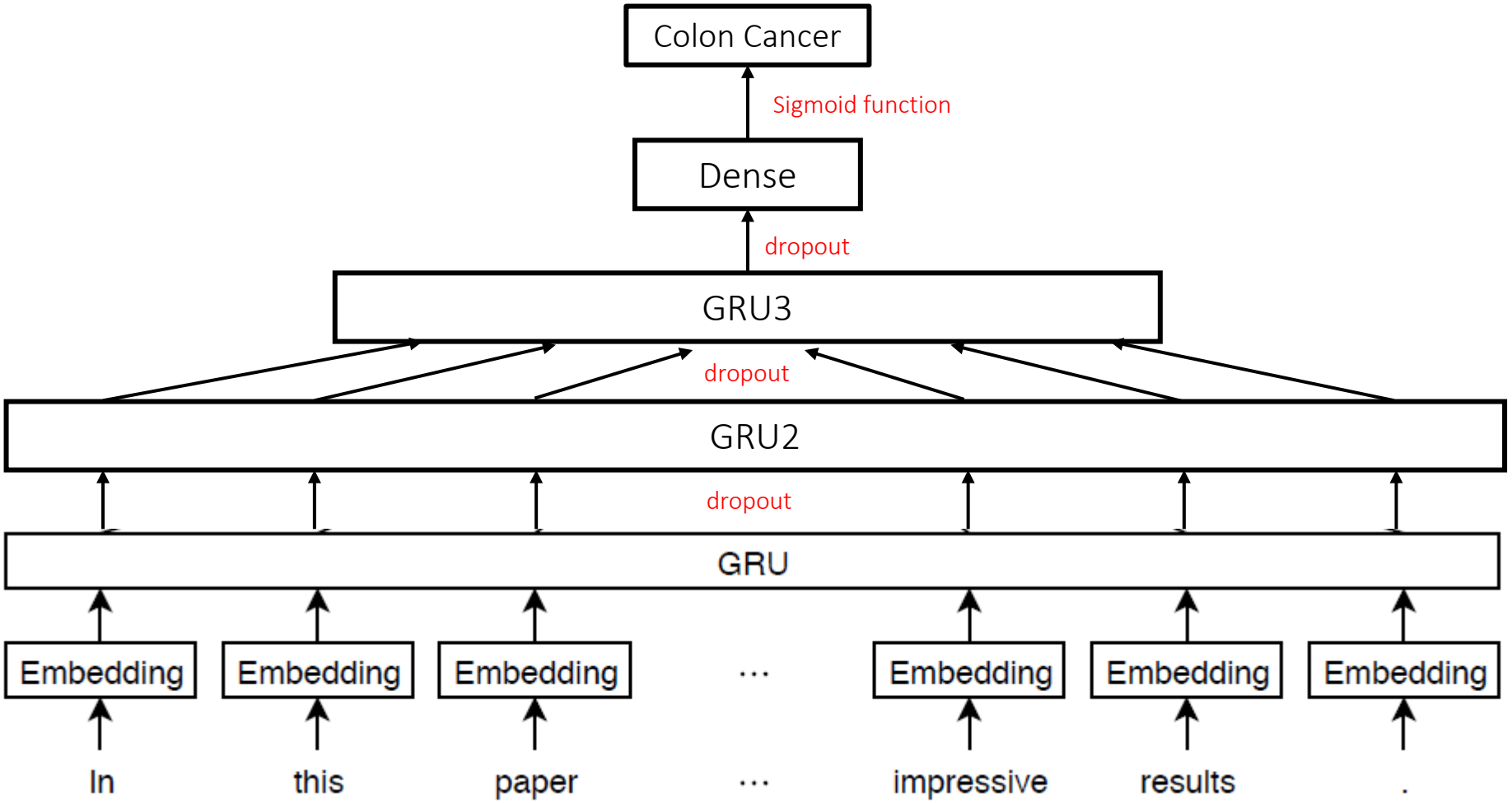


Model Development

Two models were applied

1. Gated Recurrent Unit (GRU)
   - Pre-trained GloVe (Global Vectors for Word Representation)

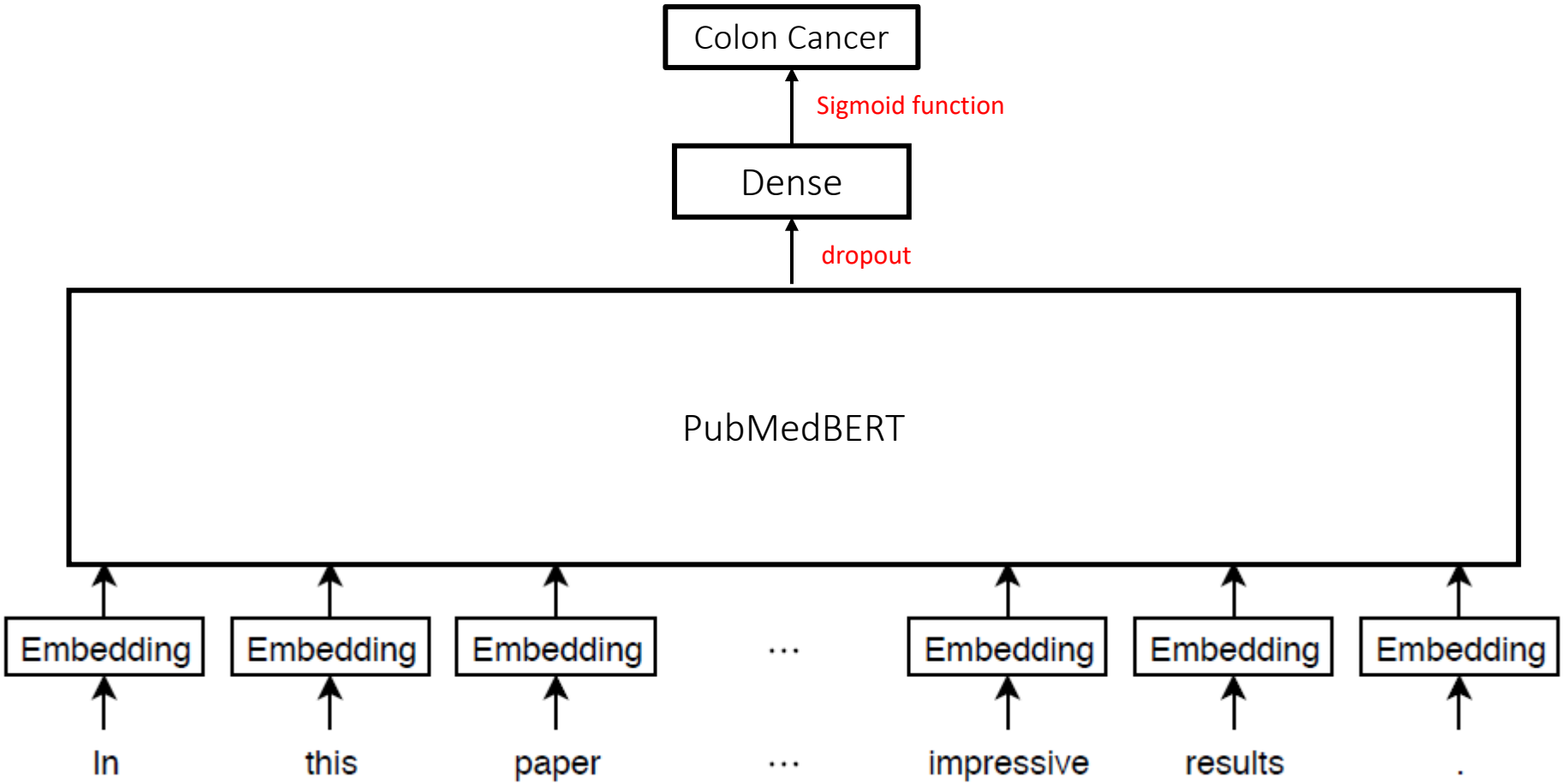
2. Bidirectional Encoder Representations from Transformers (BERT)
   - Pre-trained BERT on a medical domain (PubMedBERT)



https://www.mathworks.com/help/deeplearning/ug/multilabel-text-classification-using-deep-learning.html

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 241, 50)	253700
batch_normalization (BatchNormalization)	(None, 241, 50)	200
dropout (Dropout)	(None, 241, 50)	0
gru (GRU)	(None, 241, 32)	8064
dropout_1 (Dropout)	(None, 241, 32)	0
gru_1 (GRU)	(None, 241, 32)	6336
dropout_2 (Dropout)	(None, 241, 32)	0
gru_2 (GRU)	(None, 16)	2400
dropout_3 (Dropout)	(None, 16)	0
dense (Dense)	(None, 1)	17

Total params: 270,717
Trainable params: 16,917
Non-trainable params: 253,800



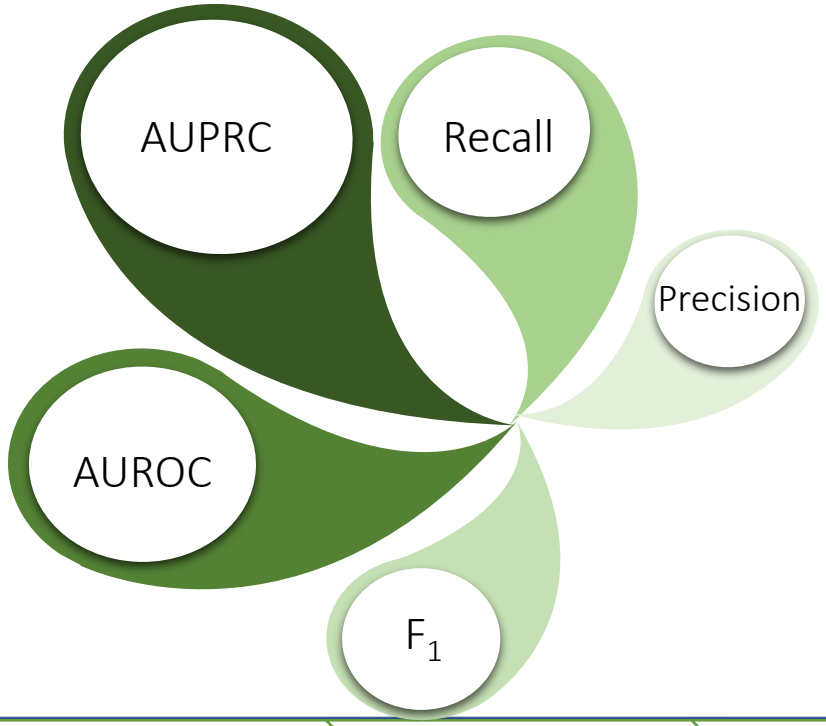
https://www.mathworks.com/help/deeplearning/ug/multilabel-text-classification-using-deep-learning.html

Layer (type)	Output Shape	Param #	Connected to
text (InputLayer)	[(None,)]	0	[]
keras_layer (KerasLayer)	{'input_mask': (None, 128), 'input_word_ids': (None, 128), 'input_type_ids': (None, 128)}	0	['text[0][0]']
keras_layer_1 (KerasLayer)	{'sequence_output': (None, 128, 768), 'encoder_outputs': [(None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768)], 'pooled_output': (None, 768), 'default': (None, 768)}	109482241	['keras_layer[0][0]', 'keras_layer[0][1]', 'keras_layer[0][2]']
dropout (Dropout)	(None, 768)	0	['keras_layer_1[0][13]']
output (Dense)	(None, 1)	769	['dropout[0][0]']

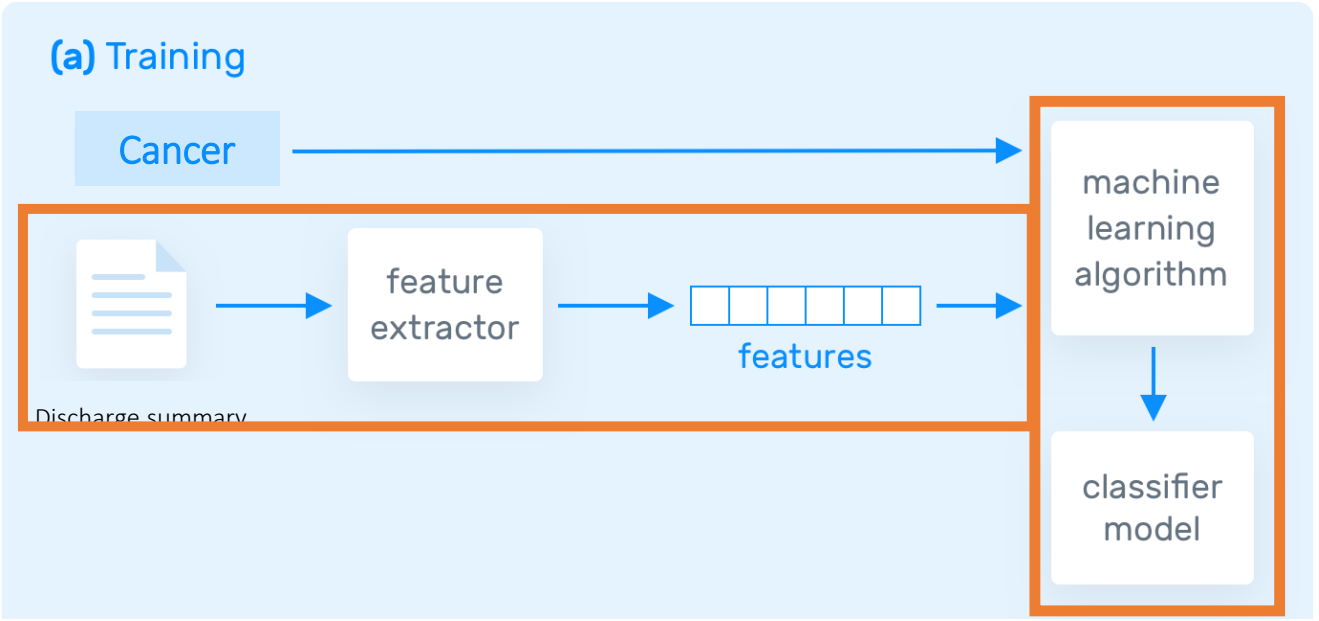
Total params: 109,483,010
Trainable params: 769
Non-trainable params: 109,482,241

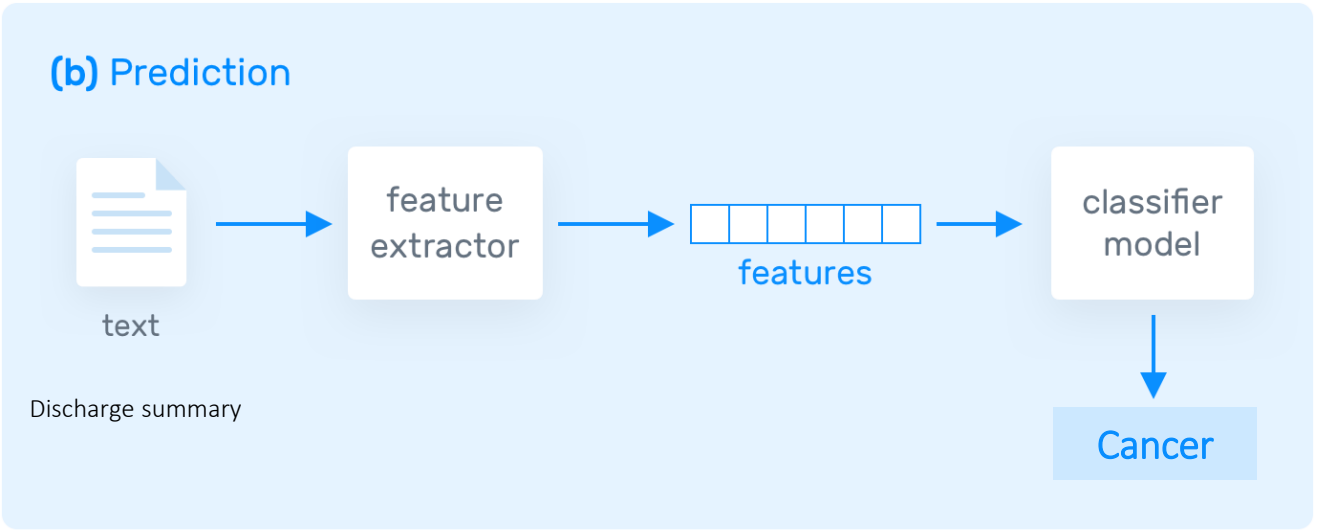
Model evaluation

Evaluation metrics

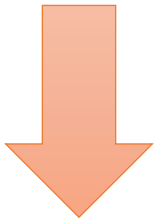


Monitor model overfitting by the percent difference between training and test performances


(a) Training
Cancer
feature extractor
features
machine learning algorithm
classifier model
Discharge summary


(b) Prediction
feature extractor
features
classifier model
text
Discharge summary
Cancer

Let's get your hands dirty



shorturl.at/mqI24