Tell Me What To Do: Prioritizing Data Labeling for NLP Systems with Active Learning

Andre Goncalves

Research Scientist – Machine Learning

Open Data Science Conference West

November 18, 2021

LLNL-PRES-829265

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Outline

- Background
- NLP systems for medical applications
- Active Learning
- Cancer pathology report classification
- Final thoughts

Little bit about myself

- Research Scientist within the Machine Learning group at LLNL since 2017
- Machine Learning interests:
  - Transfer and multitask learning
  - Probabilistic deep learning models
  - Uncertainty quantification in large ML models
  - Optimization for ML
- Recent projects:
  - ML for healthcare: cancer prognostics with MTL and synthetic EHR data generation
  - Deep Learning for climate (sub-)seasonal forecasting
  - Deep Reinforcement Learning for antibody design
  - ML for microbiome profile characterization

Background

Background

- Project developed with the National Cancer Institute (NCI) and other national laboratories:

NIH NATIONAL CANCER INSTITUTE

Frederick National Laboratory

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY

OAK RIDGE National Laboratory

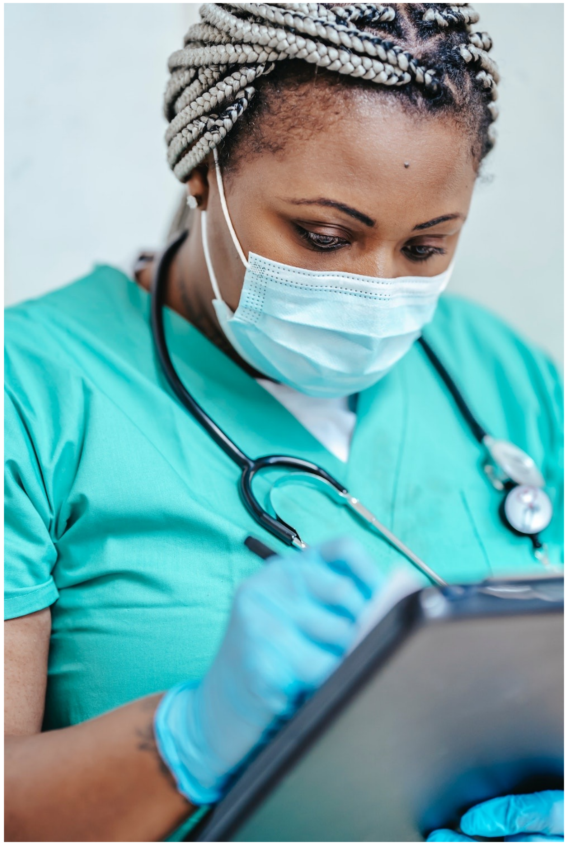
Argonne NATIONAL LABORATORY

- Development of natural language processing and deep learning algorithms to population-based cancer statistics collected by NCI's SEER program.
- Automate pathology reports classification and annotation process.

NLP Systems for Medical Applications

NLP systems for medical applications

- Many of the medical procedures, exams, and treatments are stored in written format.
- Notes contain the state and evolution of the patient throughout the treatment/medical procedure.
- High volumes of unstructured reports go to electronic health records (EHR) systems daily.
- 80% of healthcare-associated text data is unstructured and goes largely unutilized.[1]



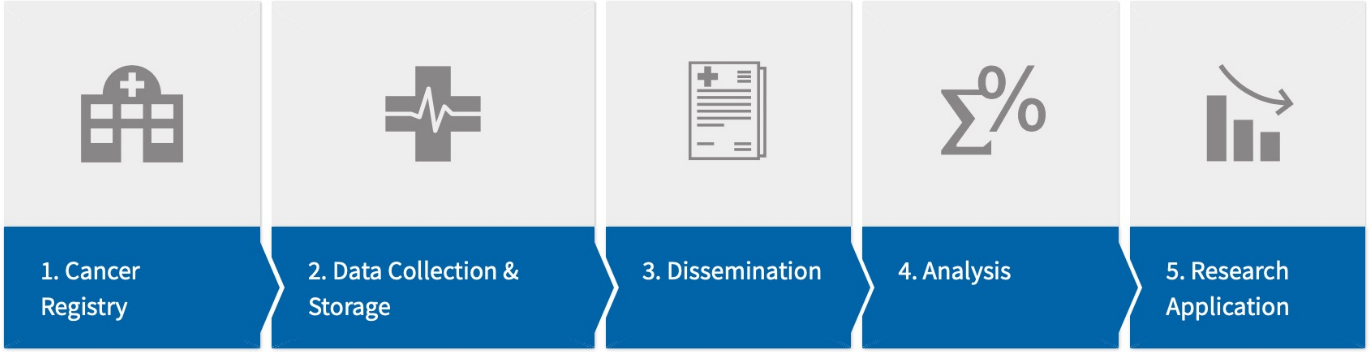
[1] https://www.foreseemed.com/natural-language-processing-in-healthcare

NLP systems for medical applications

- Physicians don’t all “speak the same way” and there is not always a diagnosis consensus.
- This pile of data contains information that can dramatically improve current understandings of treatments and medical protocols.

NLP systems in Cancer Surveillance

- In the US, NCI's *Surveillance, Epidemiology, and End Results* (SEER) program is responsible for collecting, processing, and aggregating cancer data through local **Cancer Registries**.
- Surveillance data serve as a foundation for cancer research and are used to **plan and evaluate cancer prevention and control interventions**.

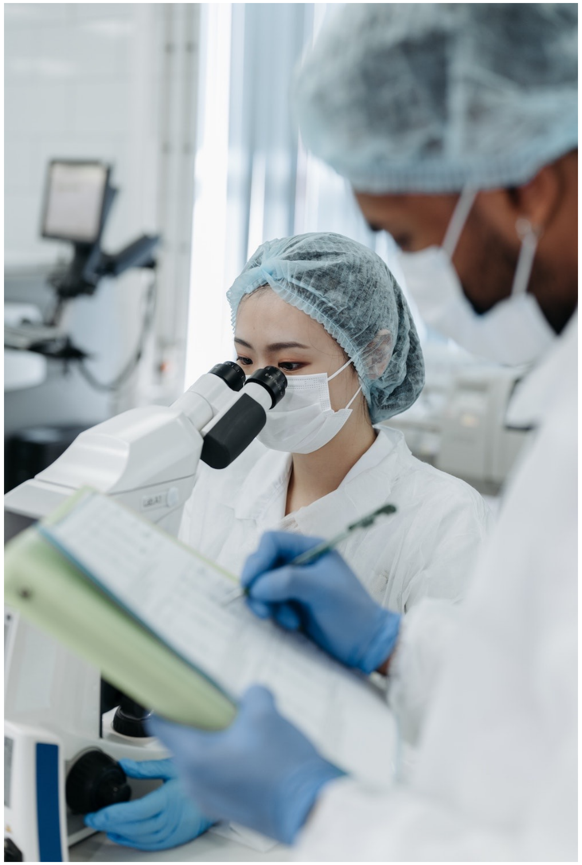


Source: [1]

[1] https://seer.cancer.gov/registries/cancer_registry/index.html

NLP systems in Cancer Surveillance

- Critical cancer data elements such as primary tumor site and tumor morphology reside in unstructured **cancer pathology reports** that are written during the time of diagnosis.
- Such reports are not only ungrammatical, fragmented, and marred with typos and abbreviations, but also exhibit linguistic variability across pathologists.



NLP systems in Cancer Surveillance

- Pathology reports:
  - "A pathology report is a document that contains the diagnosis determined by examining cells and tissues under a microscope. The report may also contain information about the size, shape, and appearance of a specimen as it looks to the naked eye." [1]

- Pathologist is a doctor responsible for examining and writing a pathology report.

- Pathology reports are essential for cancer diagnosis and staging, therefore helping to determine treatment options.

[1] https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet

NLP systems in Cancer Surveillance

- Show pathology report examples:
  - Example 1
  - Example 2

NLP systems in Cancer Surveillance

- Cancer information is **manually** extracted from the pathology reports by **Certified Tumor Registrars** (CTRs).

- Cancer registries face challenges scaling the manual effort to handle the increasing volumes of clinical reports they need to process and the amount of information they need to capture per report.

- NCI’s effort to **automate** some of the involved processes **using NLP systems**.

Challenges in Pathology Report Classification

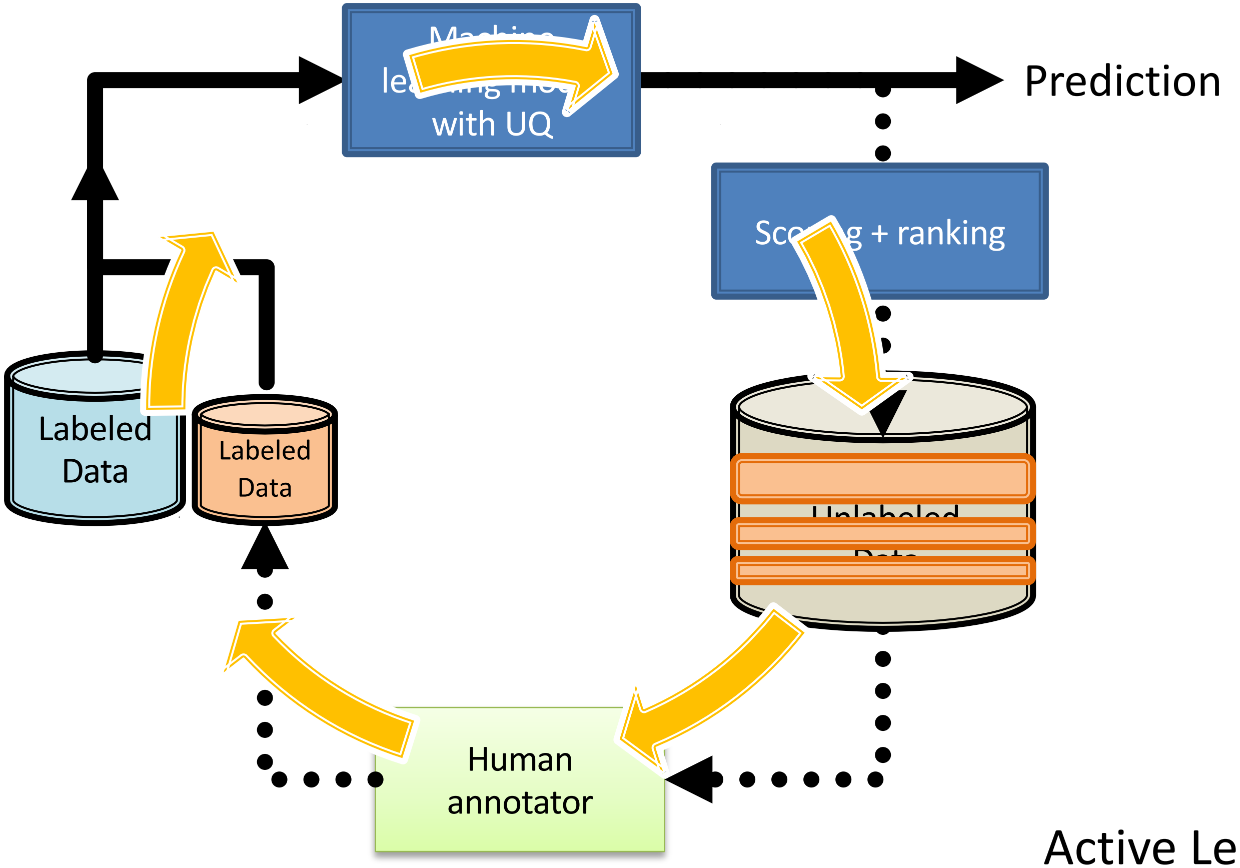
- Reports lack of universal structure, variation in linguistic patterns, and the presence of medical jargons.
- Reports are long documents (several pages) but only a few keywords may be relevant to a specific classification task.
- Long-distance linguistic dependencies across different sections of a document.
- Certain tasks may contain several classes (i.e., 525 histology types), and the number of documents per class tends to be highly unbalanced.

NLP for Pathology Report Classification

- State-of-the-art data-driven NLP systems require large sets of labeled data to reach reasonable accuracy.
- Few skilled Certified Tumor Registrars are available to provide labels to existing pathology reports.
- Solution: **use active learning to direct which pathology reports to label first, so the performance of NLP is maximized.**

Active Learning

Active Learning



Active Learning Loop

Active Learning

- Active learning (AL) **seeks to maximize an ML model's performance gain while labeling the fewest number of samples possible** (reduce annotation cost)**.**

- AL can be seen as a way to integrate humans into the machine learning process (**human-in-the-loop** paradigm).

- AL is a type of **semi-supervised learning**, meaning models are trained using both labeled and unlabeled data.

Active Learning

- Keys components in Active Learning setup:
  - — Labeled data << unlabeled data
  - — Acquisition function (*scoring function*):
    - Provides a measure of *informativeness* of a given sample to the ML model
    - Indicates how helpful that sample would be to the classifier/regressor if it were to be trained on that sample
- Active Learning loop can be executed indefinitely or until a target performance is reached or the budget is depleted.

AL categorization: Query mechanisms

- Membership query synthesis:
  - learner can request to query the label of any unlabeled sample in the input space, including the sample generated by the learner.
- Stream-based selective sampling:
  - makes an independent judgment on whether each sample in the data stream needs to query the labels of unlabeled samples.
- Pool-based selective sampling :
  - chooses the best query sample based on the evaluation and ranking of the entire dataset.

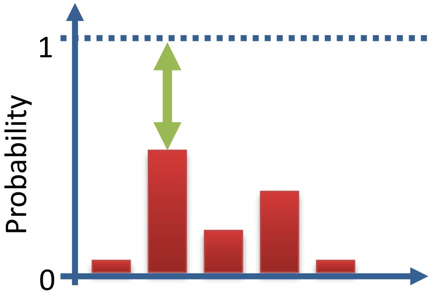
AL categorization: Scoring functions

- Uncertainty-based approach:
  - Select samples the current model are most uncertain about.
- Diversity-based approach:
  - Select samples following the underlying data distribution.
- Expected model change:
  - Calculate expected change in the model given the addition of a sample to the training data.
- Hybrid approaches:
  - Combinations of the approaches above.

Acquisition functions examples

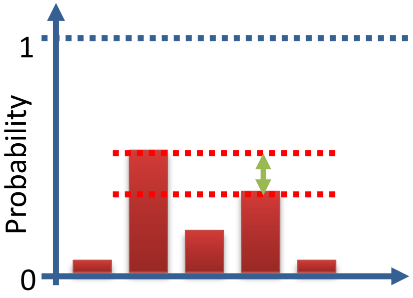
- Common uncertainty-based AF for classification:

Least Confidence



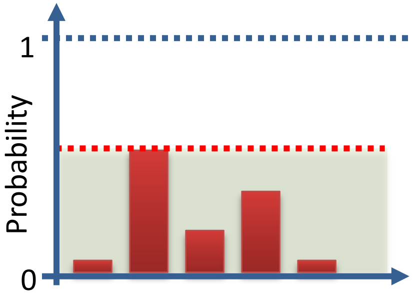
selects the instance for which it has the least confidence in its **most** likely label

Margin Sampling



selects the instance that has the smallest **difference** between the **first** and **second** most probable labels

Shannon's Entropy



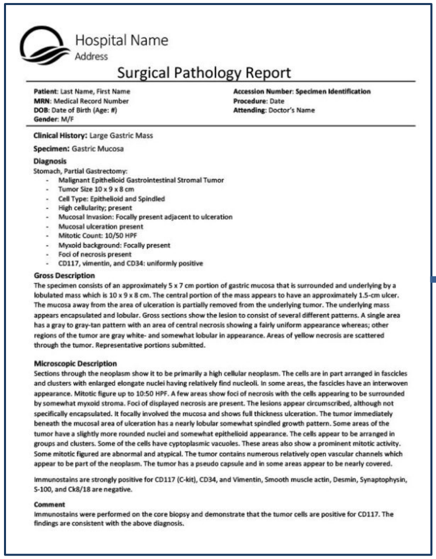
$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Experimental Study:

Cancer Pathology Report Classification

Cancer Pathology Reports Classification

- Multi-label classification problem:



Hospital Name
Address

Surgical Pathology Report

Patient: Last Name, First Name
MRN: Medical Record Number
DOB: Date of Birth (Age: #)
Gender: M/F

Accession Number: Specimen Identification
Procedure: Date
Attending: Doctor's Name

Clinical History: Large Gastric Mass

Specimen: Gastric Mucosa

Diagnosis

Stomach, Partial Gastrectomy:

- Malignant Epithelioid Gastrointestinal Stromal Tumor
- Tumor Size 10 x 9 x 8 cm
- Cell Type: Epithelioid and Spindled
- High cellularity; present
- Mucosal Invasion: Focally present adjacent to ulceration
- Mucosal ulceration present
- Mitotic Count: 10/50 HPF
- Myxoid background: Focally present
- Foci of necrosis present
- CD117, vimentin, and CD34: uniformly positive

Gross Description

The specimen consists of an approximately 5 x 7 cm portion of gastric mucosa that is surrounded and underlying by a lobulated mass which is 10 x 9 x 8 cm. The central portion of the mass appears to have an approximately 1.5-cm ulcer. The mucosa away from the area of ulceration is partially removed from the underlying tumor. The underlying mass appears encapsulated and lobular. Gross sections show the lesion to consist of several different patterns. A single area has a gray to gray-tan pattern with an area of central necrosis showing a fairly uniform appearance whereas; other regions of the tumor are gray white- and somewhat lobular in appearance. Areas of yellow necrosis are scattered through the tumor. Representative portions submitted.

Microscopic Description

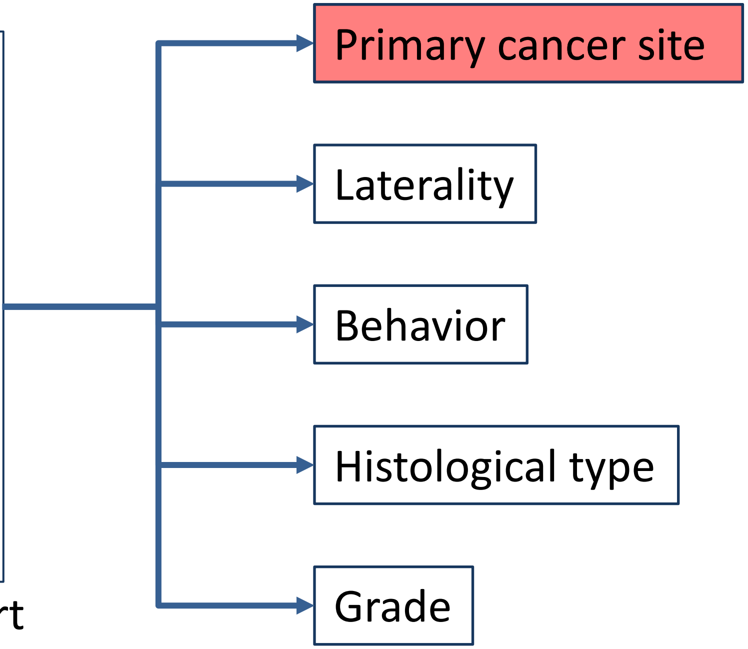
Sections through the neoplasm show it to be primarily a high cellular neoplasm. The cells are in part arranged in fascicles and clusters with enlarged elongate nuclei having relatively find nucleoli. In some areas, the fascicles have an interwoven appearance. Mitotic figure up to 10:50 HPF. A few areas show foci of necrosis with the cells appearing to be surrounded by somewhat myxoid stroma. Foci of displayed necrosis are present. The lesions appear circumscribed, although not specifically encapsulated. It focally involved the mucosa and shows full thickness ulceration. The tumor immediately beneath the mucosal area of ulceration has a nearly lobular somewhat spindled growth pattern. Some areas of the tumor have a slightly more rounded nuclei and somewhat epithelioid appearance. The cells appear to be arranged in groups and clusters. Some of the cells have cytoplasmic vacuoles. These areas also show a prominent mitotic activity. Some mitotic figured are abnormal and atypical. The tumor contains numerous relatively open vascular channels which appear to be part of the neoplasm. The tumor has a pseudo capsule and in some areas appear to be nearly covered.

Immunostains are strongly positive for CD117 (C-kit), CD34, and Vimentin, Smooth muscle actin, Desmin, Synaptophysin, S-100, and Ck8/18 are negative.

Comment

Immunostains were performed on the core biopsy and demonstrate that the tumor cells are positive for CD117. The findings are consistent with the above diagnosis.

Cancer Pathology report



Primary cancer site

Laterality

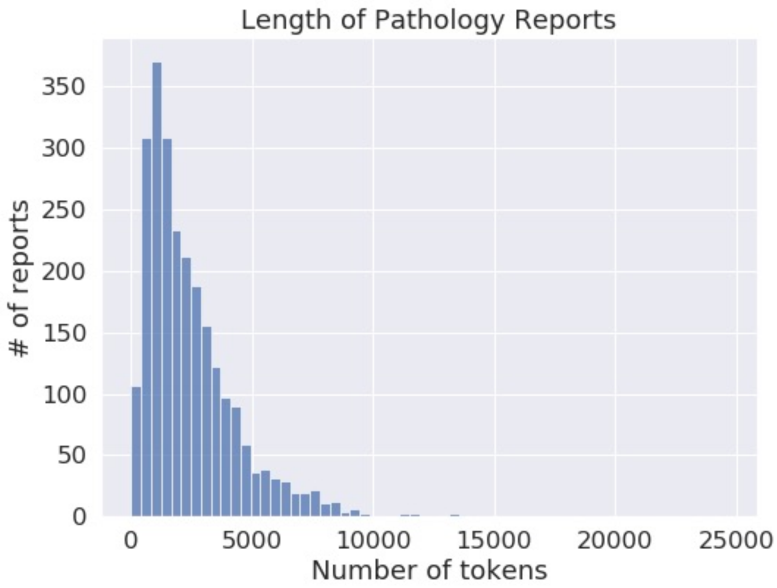
Behavior

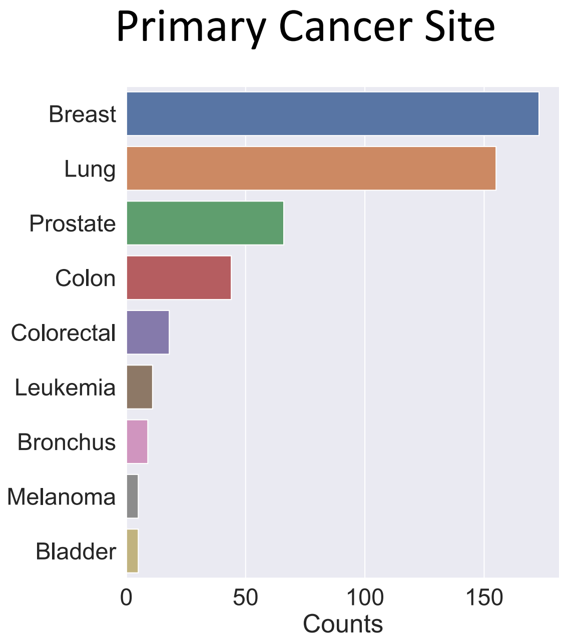
Histological type

Grade

Pathology Reports Classification: Data

- Data description:
  - — 2,500 de-identified cancer pathology reports:
    - 486 reports with **Primary Cancer Site** labelled



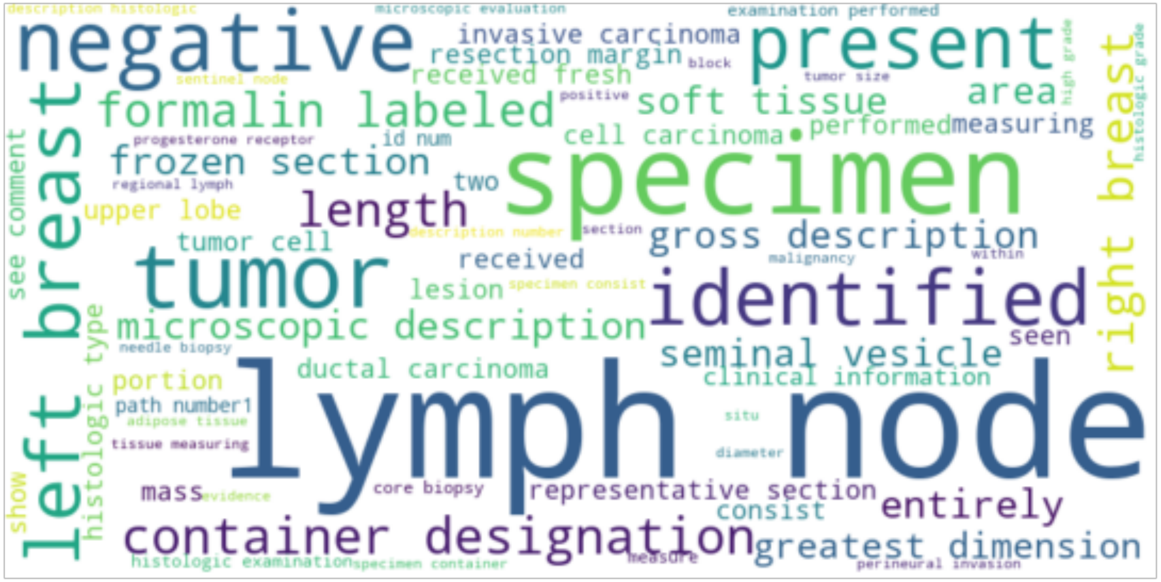


Pathology Reports Classification: Data cleaning

- Cleaning and normalization steps (ntlk library):
  - Removed tags, section markers, headers, and footers
  - Removed punctuations
  - Removed non-ascii characters
  - Replaced integers to textual representations (1 → one, …)
  - Converted to lowercase
  - Removed single character words
  - Removed stop words
  - Manually fixed some typos

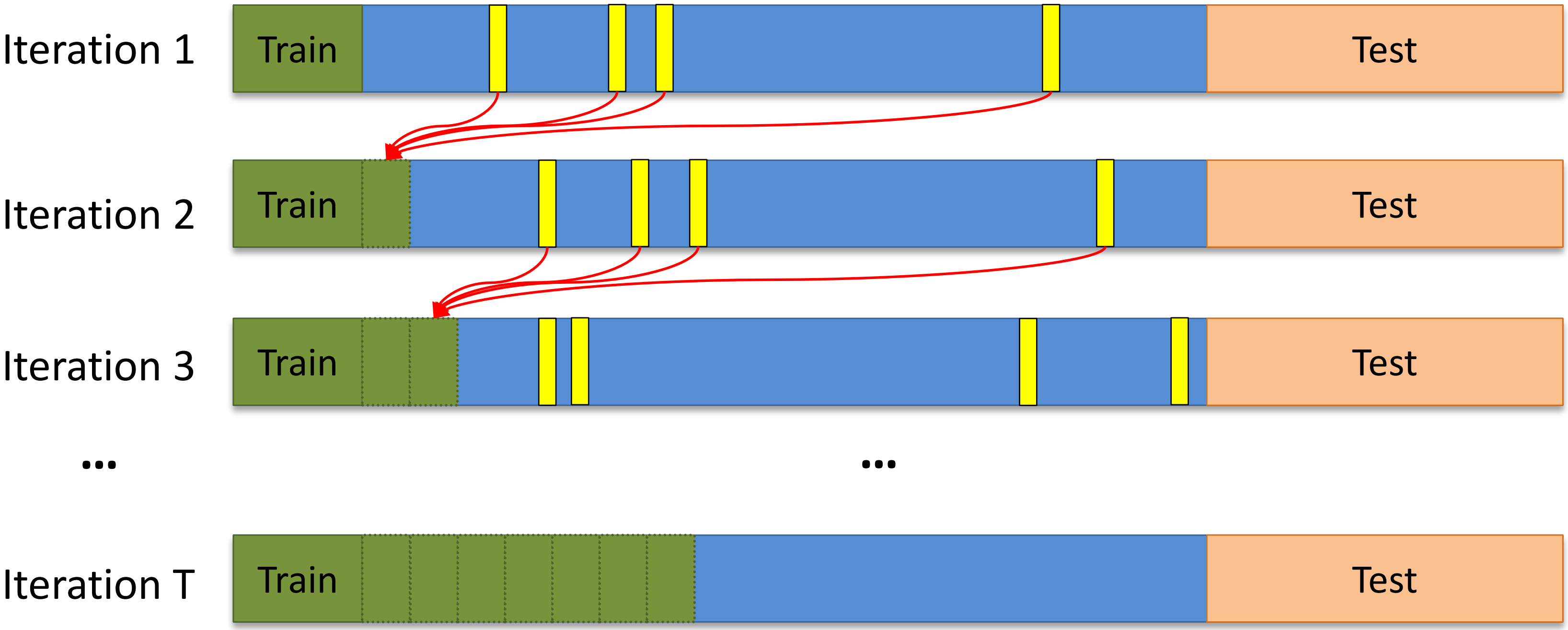
Pathology Reports Classification

- Most frequent words



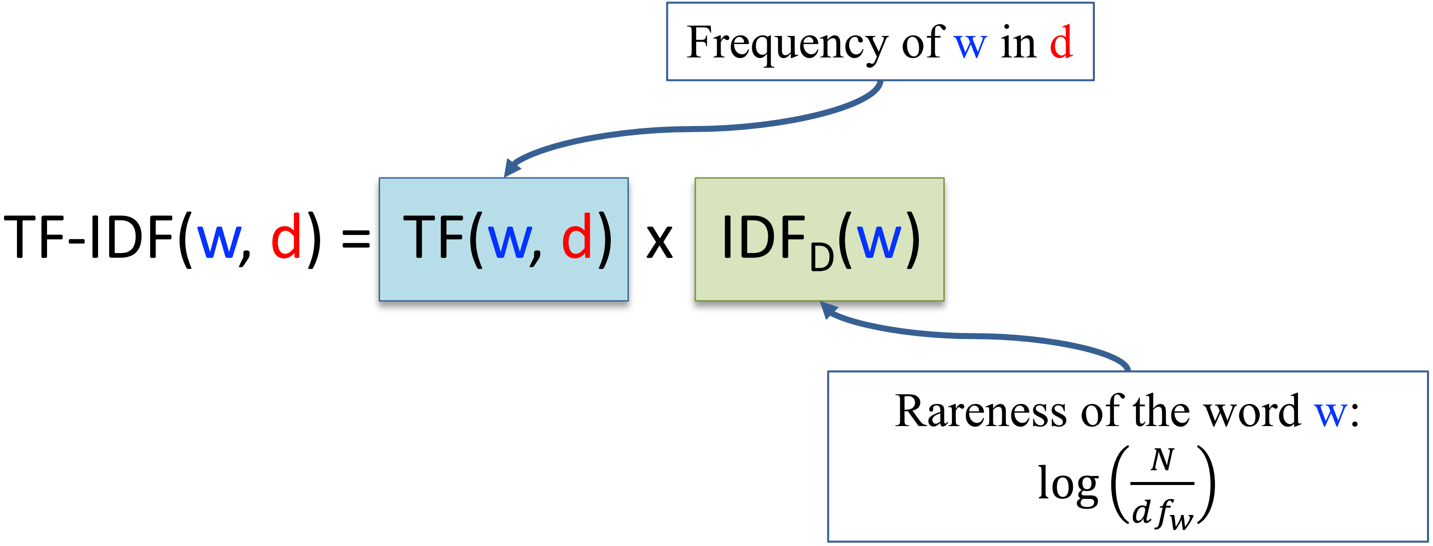
Experimental Setup

- Stratified train/test/unlabeled split to account for class imbalance



Text Feature Extraction

- TF-IDF BoW with PCA (50 dimensions)
  - Bag-of-words (BoW) concept



Frequency of w in d

$$\text{TF-IDF}(w, d) = \boxed{\text{TF}(w, d)} \times \boxed{\text{IDF}_D(w)}$$

Rareness of the word w:

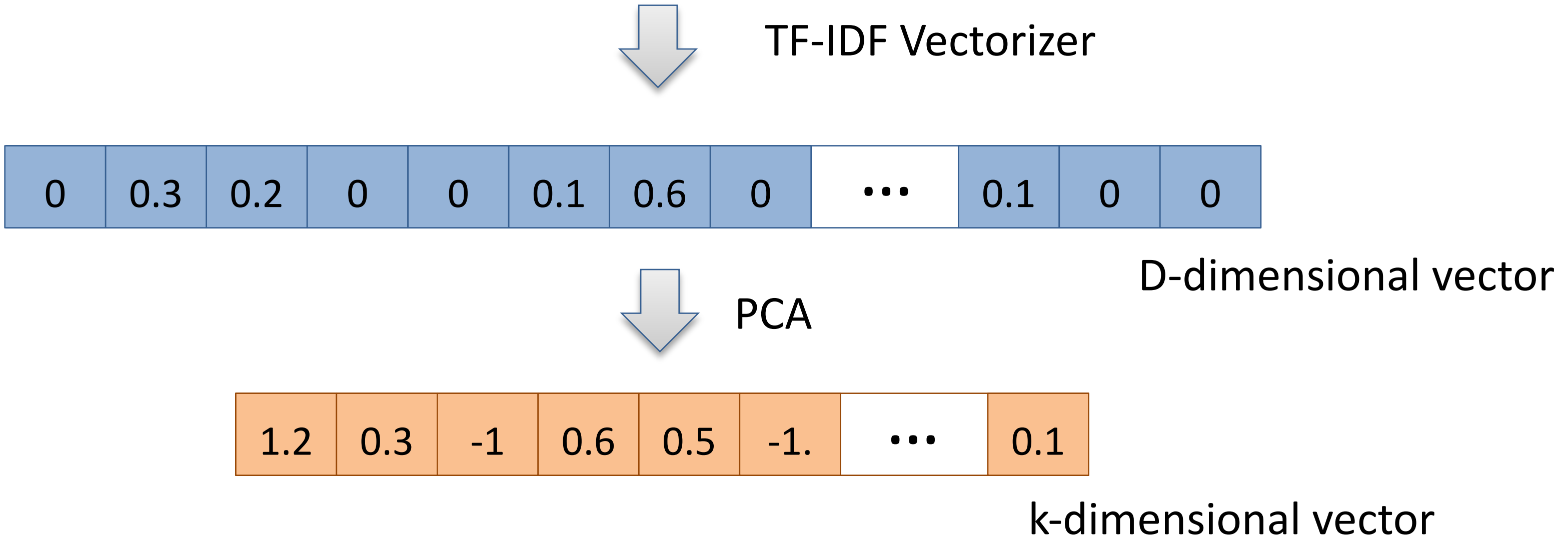
$$\log\left(\frac{N}{df_w}\right)$$

- Sklearn's TfidfVectorizer
- PCA is applied to the TF-IDF vectors

Text Feature Extraction

- TF-IDF BoW with PCA (50 dimensions)

“Microscopically there is a ragged ulcer penetrating to within 5 mm of the serosal surface. The ulcer is surrounded by dense connective response and chronic inflammatory infiltrate.”



TF-IDF Vectorizer

0	0.3	0.2	0	0	0.1	0.6	0	•••	0.1	0	0

D-dimensional vector

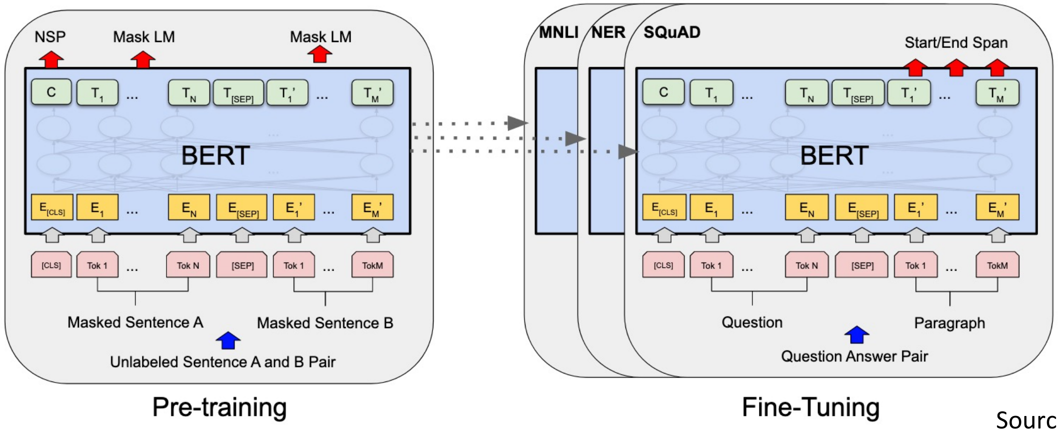
PCA

1.2	0.3	-1	0.6	0.5	-1.	•••	0.1

k-dimensional vector

Text Feature Extraction

- BERT (pre-trained Transformer-based model)



Pre-training

Fine-Tuning

Source [1]

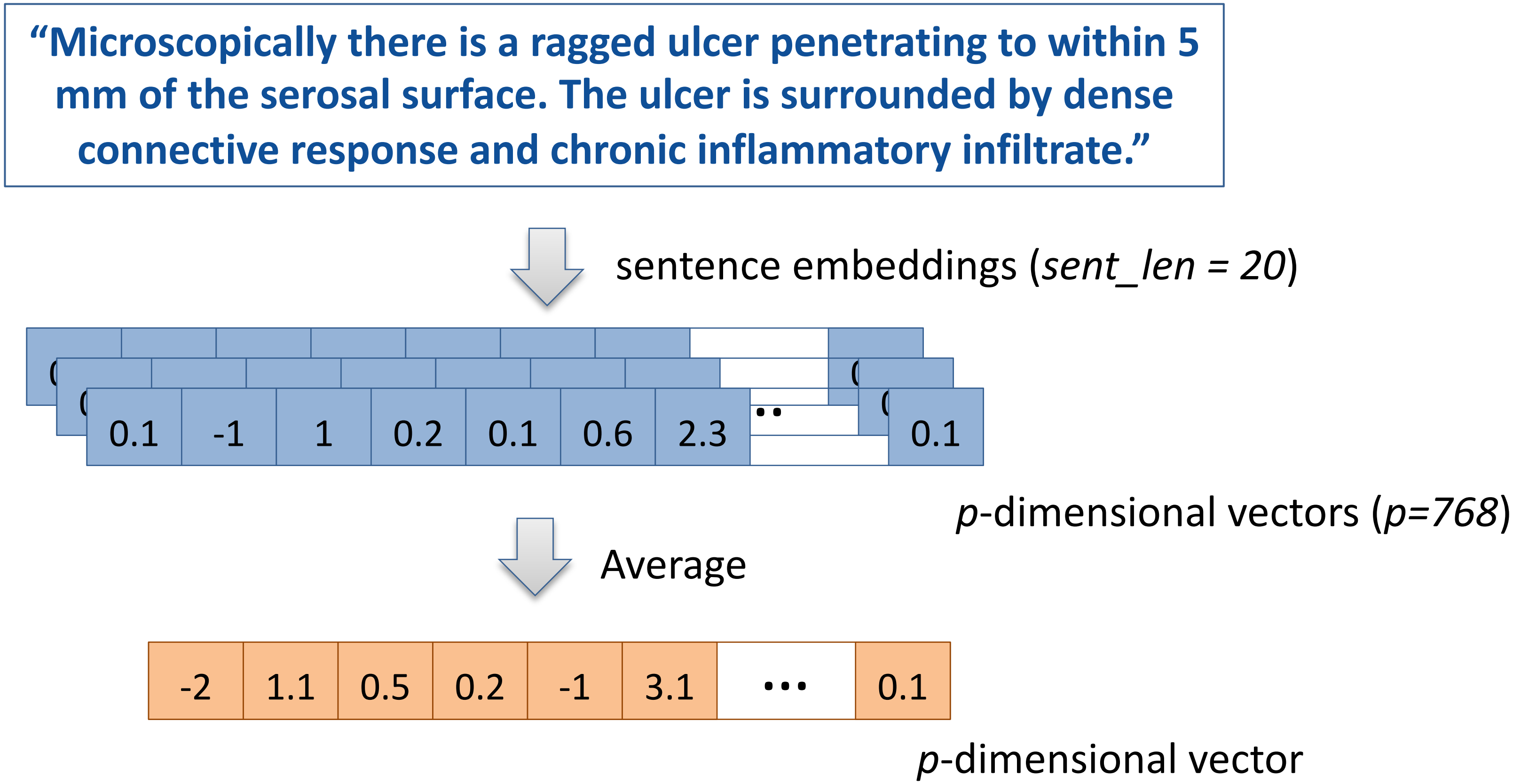
- Pathology report embedding is the mean value from all sentence's embeddings.

- Hugging-face's Transformer library (BertTokenizer, BertModel)

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Text Feature Extraction

- BERT (pre-trained Transformer-based model)



“Microscopically there is a ragged ulcer penetrating to within 5 mm of the serosal surface. The ulcer is surrounded by dense connective response and chronic inflammatory infiltrate.”

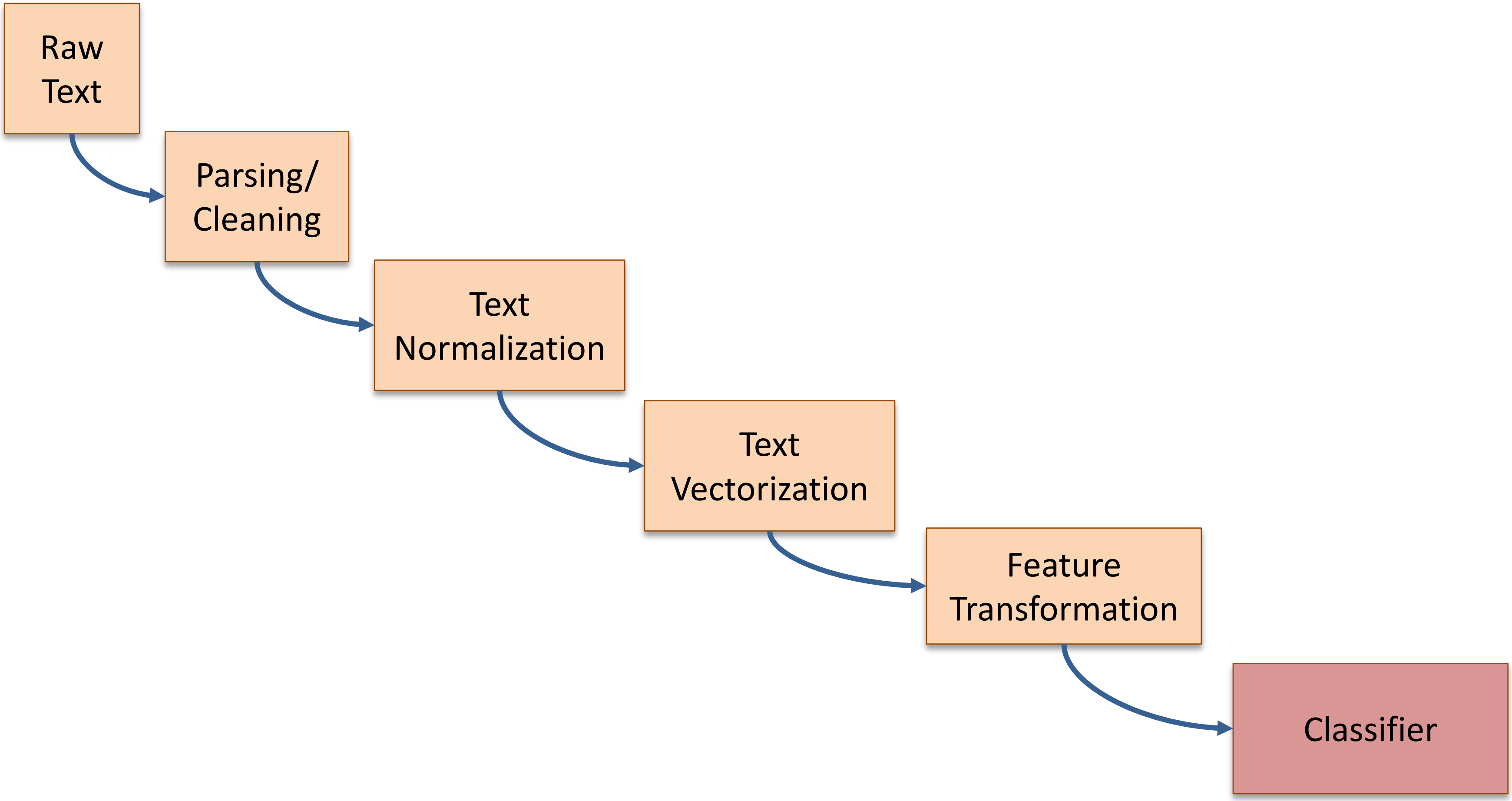
sentence embeddings (*sent_len = 20*)

p-dimensional vectors (*p=768*)

Average

p-dimensional vector

Text Analysis Pipeline



Adapted from: *Applied Text Analysis with Python* by Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda (2018)

Classifiers (with UQ)

- Deterministic (use softmax probabilities)
  - Logistic regression
  - Deterministic (regular) Neural Network

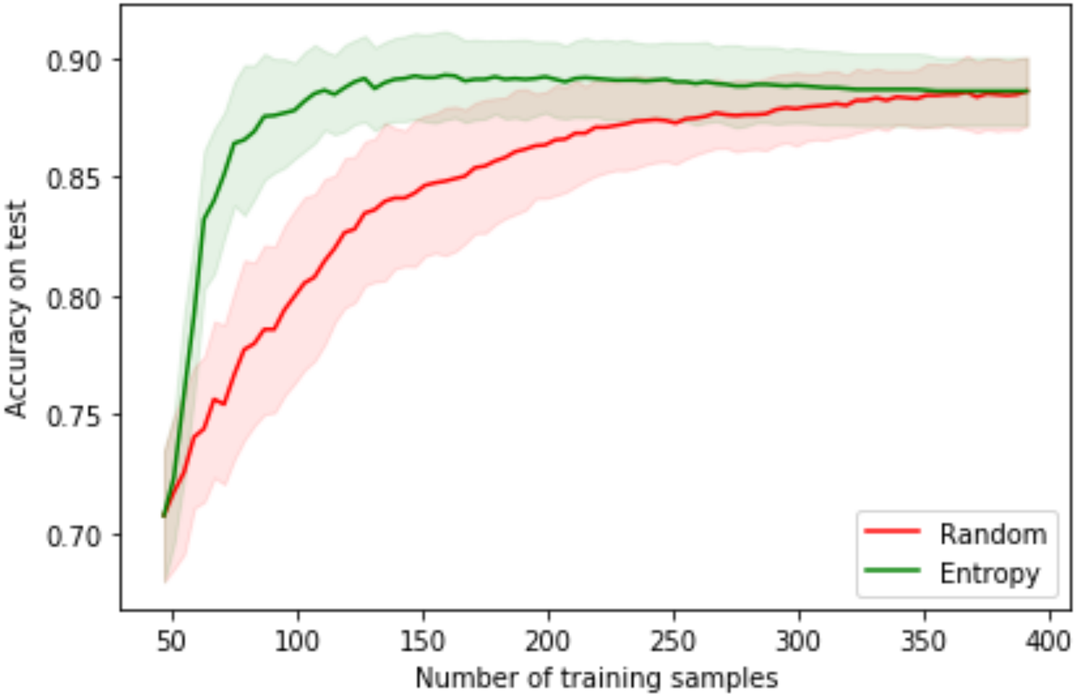
- Bayesian Models
  - Gaussian Process Classifier
  - Bayesian Linear Regression
  - Bayes-by-Backprop NN model

- Bootstrapping
  - Logistic regression
  - Random Forest
  - Neural Networks

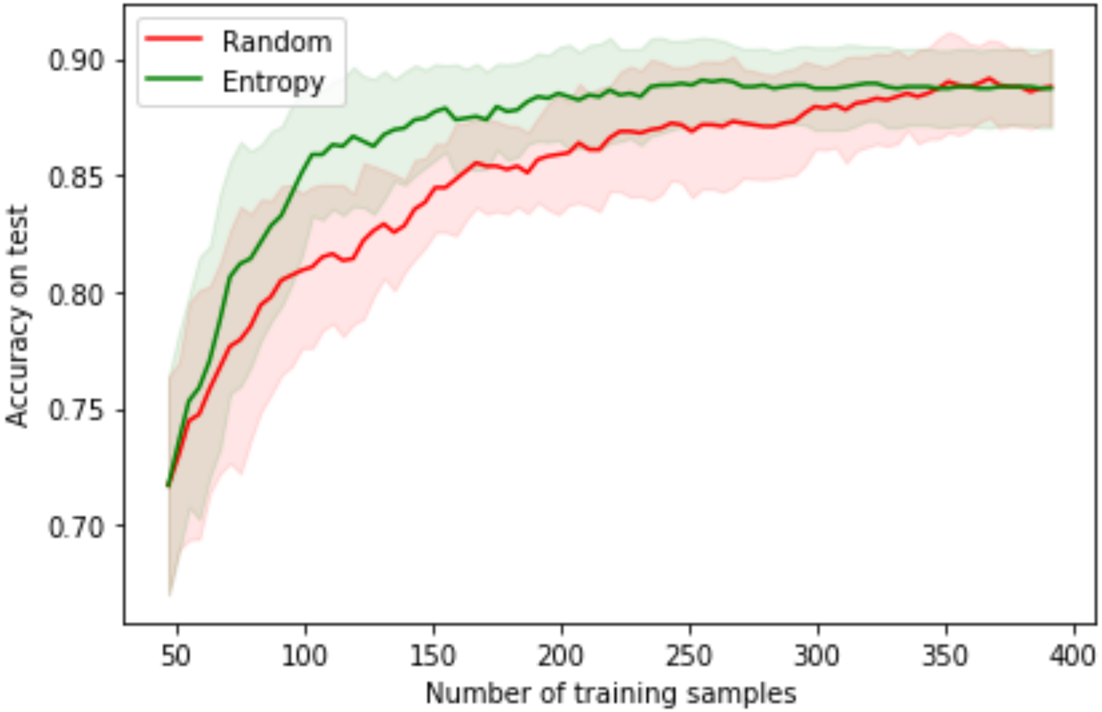
Acquisition Functions

- Random (baseline): randomly select the next set of samples to be labelled
- Entropy
- Least Confidence
- Margin Sampling

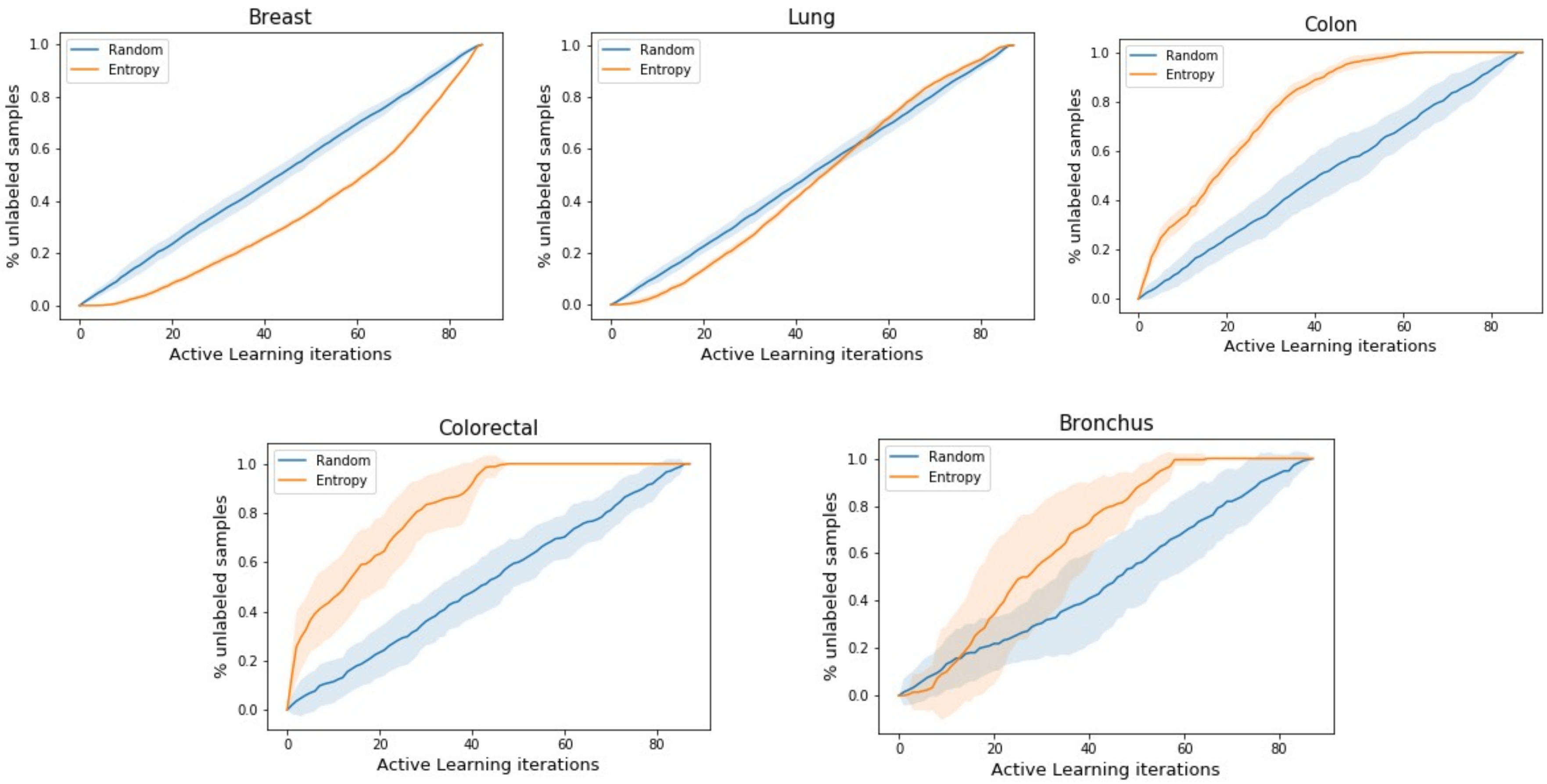
Primary Site: TF-IDF + PCA (50)



Primary Site: BERT model

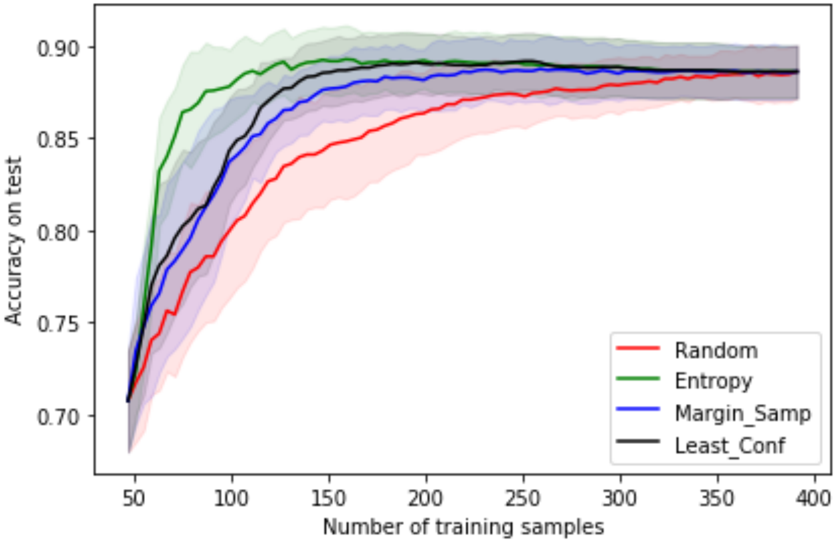


Primary Site: TF-IDF + PCA (50)

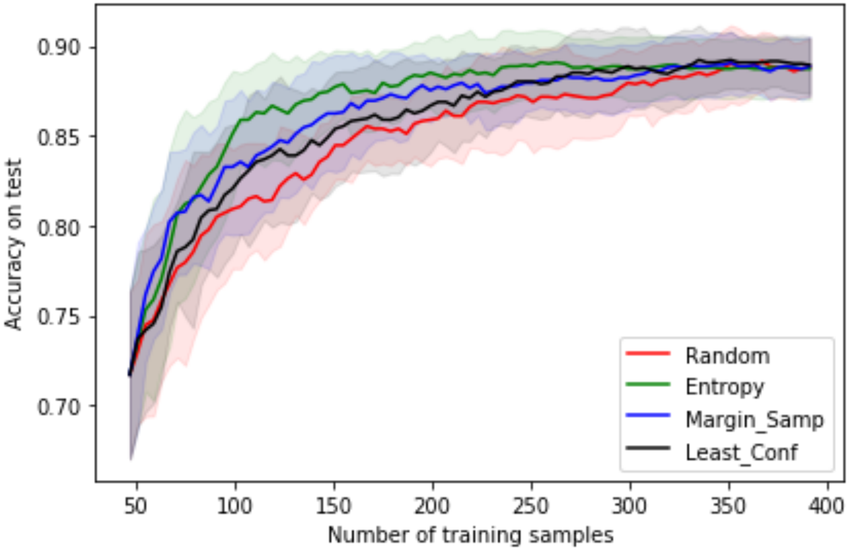


Primary Site: Other Acquisition Functions

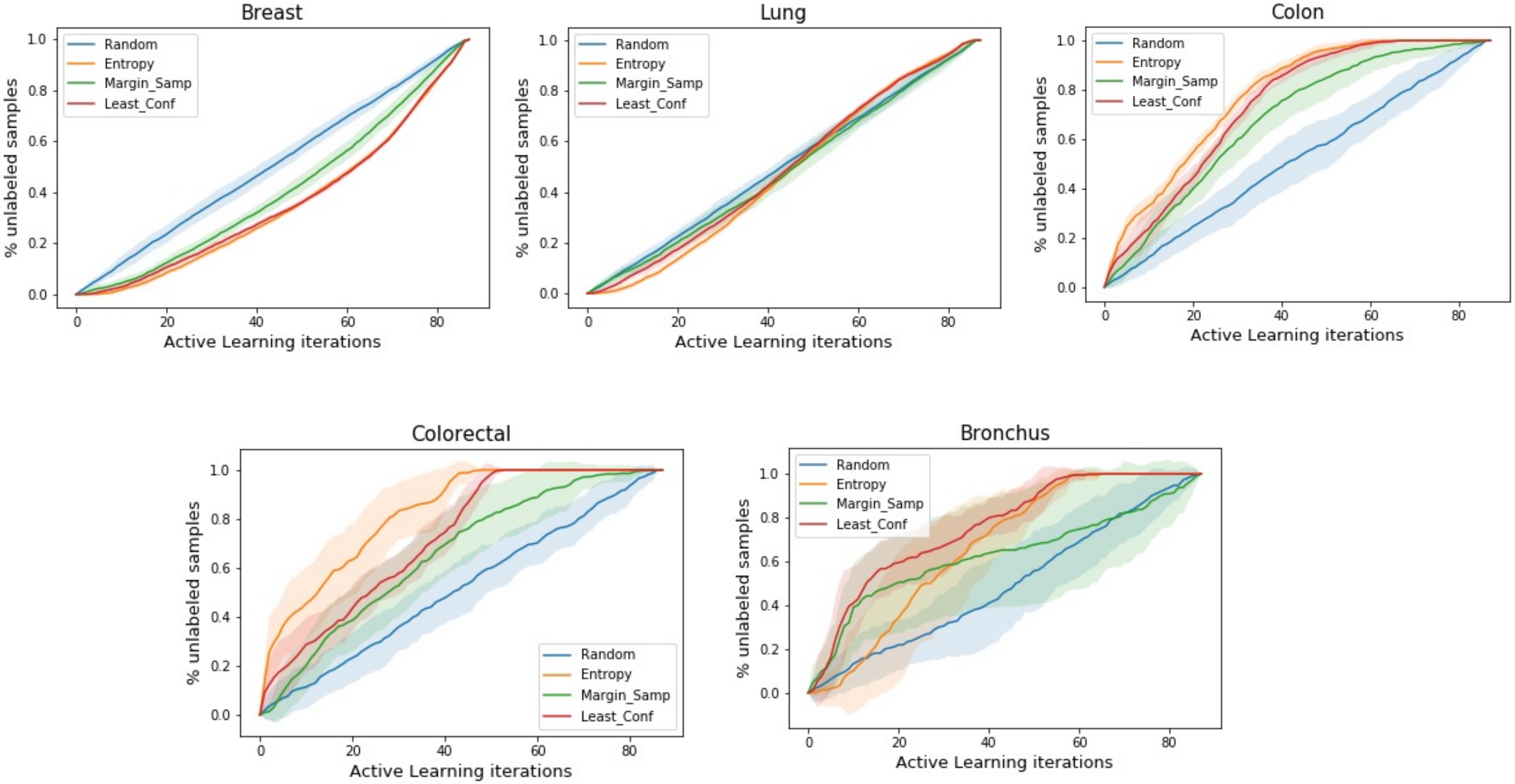
TF-IDF representation



BERT representation



Primary Site: Other Acquisition Functions



Final Thoughts

Final Thoughts

- Automated pathology report annotation can reduce the burden of manual labeling by CTRs at the cancer registries.
- It can accelerate data release and provide abundancy of labeled cancer records for statistics/machine learning research.
- NLP is a powerful tool for automating manual labeling processes.
- Active Learning helps maximize a machine learning model’s performance while reducing labeling costs.
- Active Learning is a general concept that can be used with many base ML models and setups.

Bibliography

- K. De Angeli *et al.,* "Deep active learning for classifying cancer pathology reports." *BMC bioinformatics* 22, no. 1 (2021): 1-25.

- M. Alawad *et al.*, "Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks." *Journal of the American Medical Informatics Association* 27, no. 1 (2020): 89-98.

- P. Ren *et al.*, "A survey of deep active learning." *arXiv preprint arXiv:2009.00236* (2020).

Thank you.

Andre Goncalves

andre@llnl.gov

Lawrence Livermore National Laboratory

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.