Combining Natural Language Processing with Machine-Learning Models for Detecting Progression of Disease in Radiology Report of Cancer Patients

Memorial Sloan Kettering Cancer Center

Lior Gazit, Michael Offin, Richard K. Do, Aryeh Caroline, Krishna Juluru, Marjorie G. Zauderer, Charles M. Rudin

Memorial Sloan-Kettering Cancer Center, New York, NY, USA

Motivation

- Cancer patients undergo both clinical assessments (e.g. Laboratory tests) and radiologic exams (e.g. CT, MRI imaging) to determine disease extent and response to therapies.
- Oncologists routinely incorporate radiologic reports with their clinical assessment to identify which patients have **Progression of Disease (POD)** and may require changes to their treatment.
- While specific imaging criteria define POD in clinical trials (e.g. RECIST), none exist in routine clinical practice. Thus oncologists need to **interpret the subjective phrasing of radiology reports** to assess extent of disease and recognize POD.
- In every oncologic radiology report, a summary of disease extent is communicated in the **report "Impression" section**.
- Radiologic report impressions are thus valuable data sources for automated identification of POD by computer based techniques, such as **Natural Language Processing (NLP)**.
- The purpose of this pilot study is to evaluate the potential of NLP in generating an **algorithm to automatically identify POD** from the radiologic reports in a sample lung cancer population.
- The NLP algorithm accuracy may be further evaluated and validated in other cancer patient populations for eventual clinical and research applications.

Cohort and Data

- Single institution retrospective review of **149 patients with non-small cell lung cancer**.
- Patients were selected based on the presence of sensitizing *EGFR*-mutations or rearrangements in *ALK/RET* rearrangements.
- All patients received at least one line of molecularly targeted therapy during their disease course.
- All imaging (MRI, CT, and PET) were evaluated for these patients from **September 2012 through July 2017**.
- In total, **1117 images** were reviewed.
- The median number of radiologic reports reviewed per patient was 6 (ranging 1 – 42).
- Each imaging modality resulted in a text report that **was classified into one of two classes: POD or not POD** which was assessed by a thoracic oncologist reviewing the report impressions.
- Further investigation was performed by exploring patients' medical records so as to accommodate for POD which wasn't radiology-driven. This allowed for evaluating radiologic impressions as a POD indicator.

Data Illustration

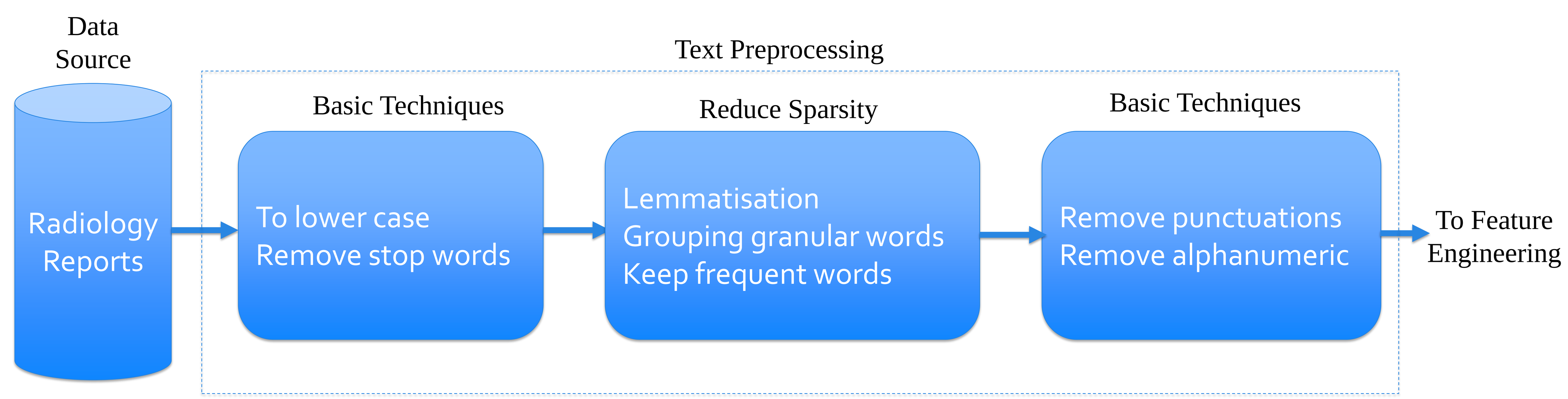
• No POD

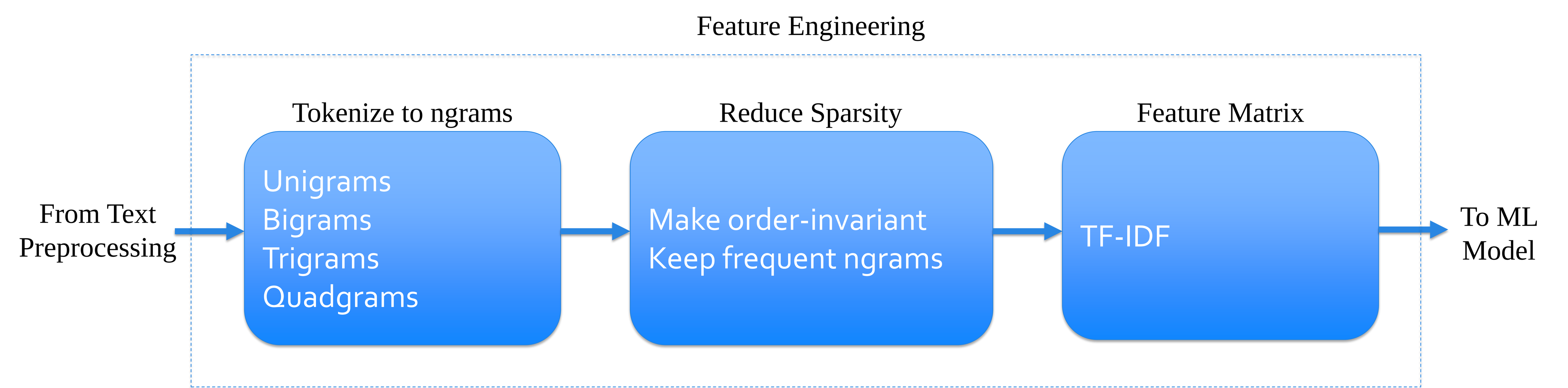
"…**no new** nodules…interval **decrease** in size of the left upper lobe lung cancer. Several additional tiny nodules within the lungs are **unchanged**…"

• POD

"…new and **enlarged** right pleural masses, consistent with malignancy. **Increased** masslike consolidation in the right upper lung, possibly…"

Text Preprocessing and Feature Engineering





Machine-Learning Model

- **Train-test split:** 80-20
- **Cross-validation scheme:** 5-fold
- **Model:** LASSO is a method of simultaneous feature selection and logistic regression.
- **Metric to optimize:** Accuracy defined as $\frac{TP+TN}{ALL}$

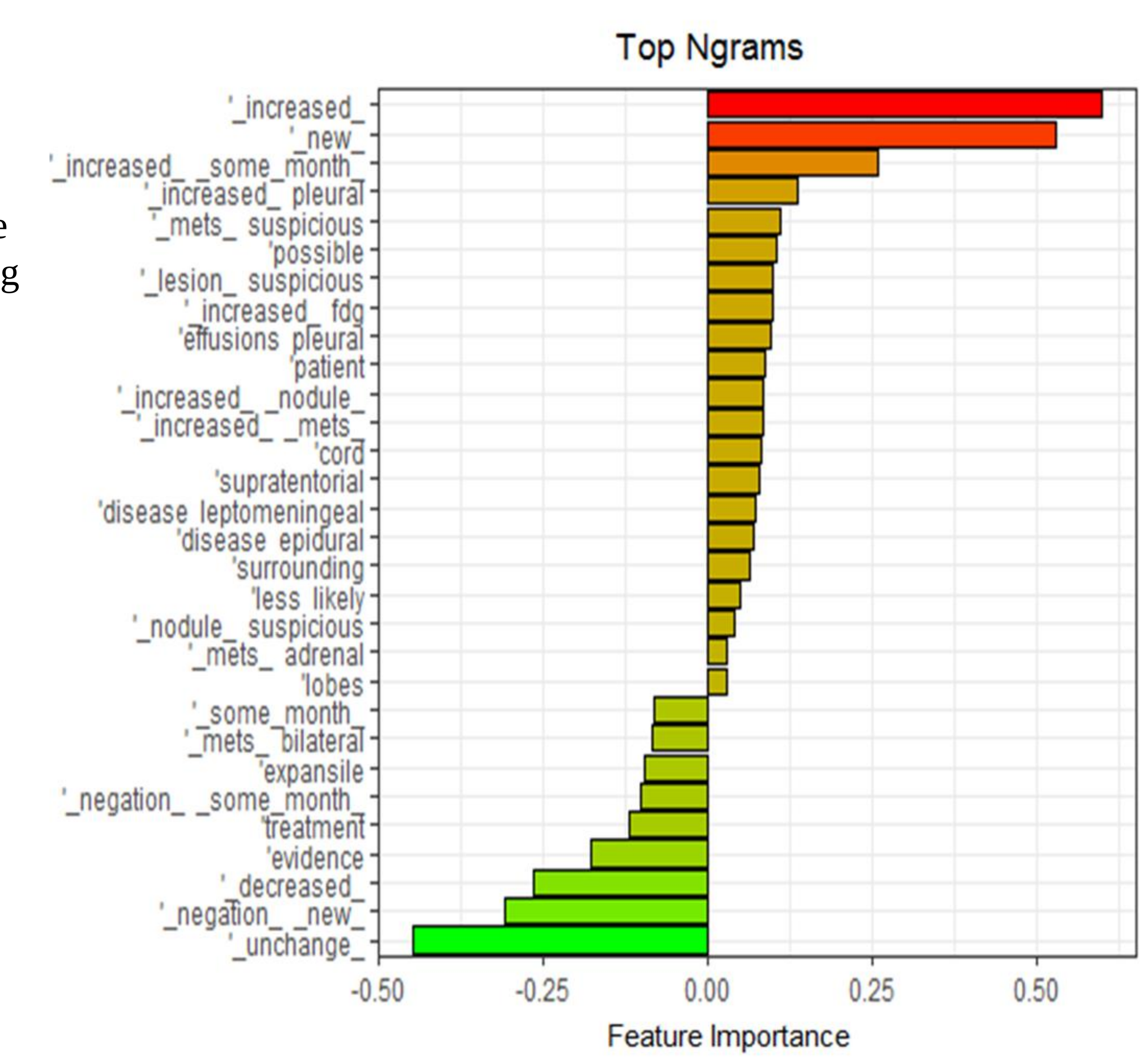
Model Goal

Natural Language Processing and machine learning (ML) algorithms have been proven successful in extracting key insights from radiology reports and translating them to structured data elements for further automated processing. In our work we implemented such algorithms to **infer clinically assessed POD from radiology reports.**

Results

$$\text{Accuracy} = \frac{TP+TN}{ALL} = \frac{13+117}{150} = 87\%$$

Total = 150	Detected: No POD	Detected: POD	
Actual: No POD	TN = 117	FP = 3	120
Actual: POD	FN = 17	TP = 13	30
	134	16	



Conclusions and Discussions

- The suggested work process and technical pipeline serve as **a computational framework to surface patient's POD status given their records of radiologic reports**.
- In order to refine the detection process and increase our accuracy we are **incorporating additional available clinical data**.