NAACCR Cancer Informatics Hackathon

Team: NLP Commanders, June 2018

Our Team

Kedar Dabhadkar

- Institute of Chemical Technology, Mumbai
- M.S. of Science in Chemical Engineering Carnegie Mellon University

Aditya Chindhade

- Birla Institute of Technology and Science 2017
- M.S of Science in Chemical Engineering Carnegie Mellon University

Dr. Jeffrey Bond

- Wisconsin Cancer Reporting System
- PhD in Biophysics, University of Rochester

Mohit Thakur

- The College of New Jersey
- MS Bioinformatics Georgia Tech University

Dr. Patrick McNeillie

- University of North Carolina 2005
- UNC Medical School 2012
- IBM Watson 2012-2017

Aakash Bhatia

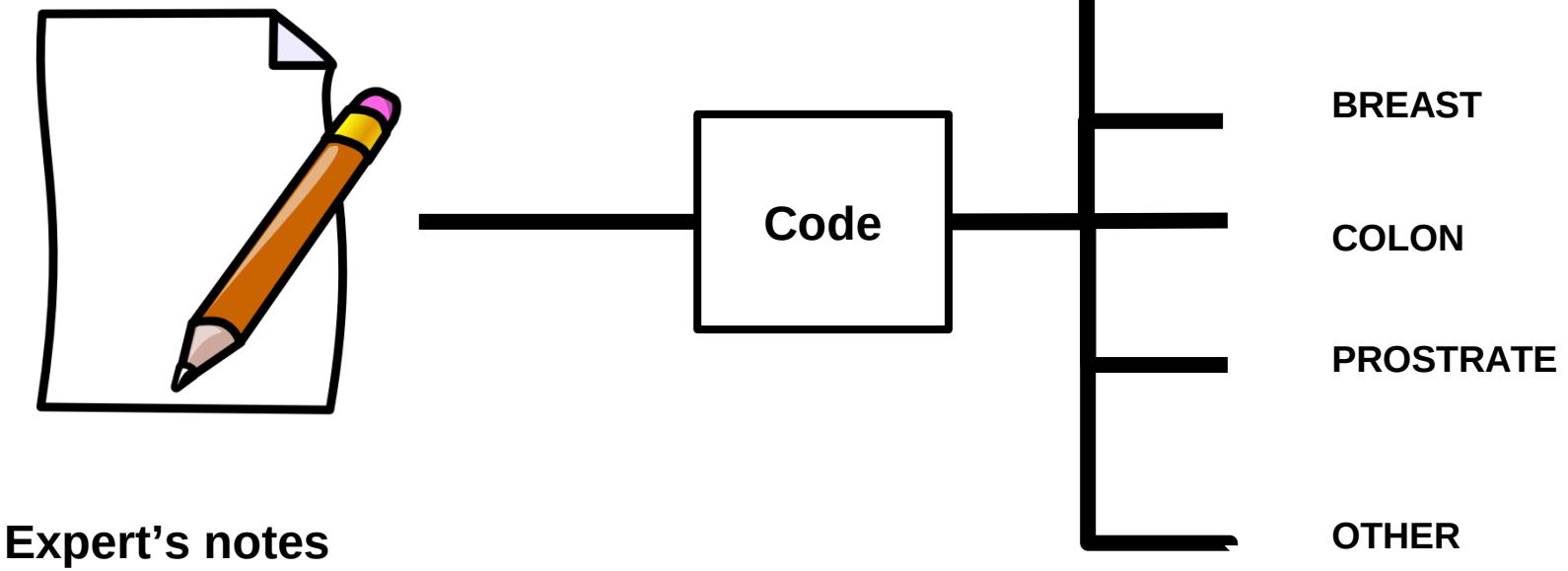
- National Institute of Technology Jaipur
- M.S. of Chemical Engineering Carnegie Mellon University

Outline

1. Challenge Introduction
2. Approach
3. Baseline Model
4. Final Model
5. Conclusion
6. Future Work

1. Challenge Introduction

Problem Statement



Unformatted Generated Sample Pathology Report (from Orchard Pathology Laboratories)

- Patient Name: Patient, John. M,Age 34 | DOB: 4/12/1979 Phone: (123) 555-1234. EMR: (123) 555-1234., PHYSICIAN INFORMATION: James Provider, MD ABC Medical 400 Royal Drive Anytown USA 12345 Phone: (123) 555-4321 Fax: (999) 555-4322., \\XOD\\REPORT DATE: 2/17/2013 TAT: [26 hours], Specimen: 2 cm polyp ascending colon 2 mm polyp in sigmoid colon Clinical History: Screening colonoscopy. Maternal hx of adenocarcinoma of colon age 57 Gross Examination A. The first container is labeled "ascending colon." It contains a polypoid piece of tan mucosal tissue measuring 2.0 cm in greatest dimension. The polyp margin is inked, sectioned, and submitted in cassettes Al and A2. B. The second container is labeled "sigmoid colon." It contains one piece of light tan mucosal tissue 0.2 cm in greatest dimension. Entirely submitted in cassette B. Microscopic Examination Microscopic Examination performed supportive of the Final Diagnosis\\XODA\\, FINAL DIAGNOSIS A. Ascending Colon SESSILE SERRATED ADENOMA (POLYP) WITH LOW-GRADE ADENOMATOUS DYSPLASIA. B. Sigmoid Colon TUBULAR ADENOMA COMMENT: \\XOD\\Patients with sessile serrated adenomas, especially with cytologic dysplasia, are at increased risk for the development of adenocarcinoma showing microsatellite instability. This progression may occur at a more rapid rate than with traditional adenomas. Complete endoscopic excision is recommended if clinically appropriate. If unresectable, repeat colonoscopy at a shortened interval (1 year), with sampling of suspicious areas or surgical resection possibly warranted.\\XODA\\ ACCESSION NUMBER 12XX0002,COLLECTION DATE: 2/15/2013 RECEIVED DATE: 2/15/2013

http://www.orchardsoft.com/files/reports/OrchardPathologyPatientReportExamples.pdf

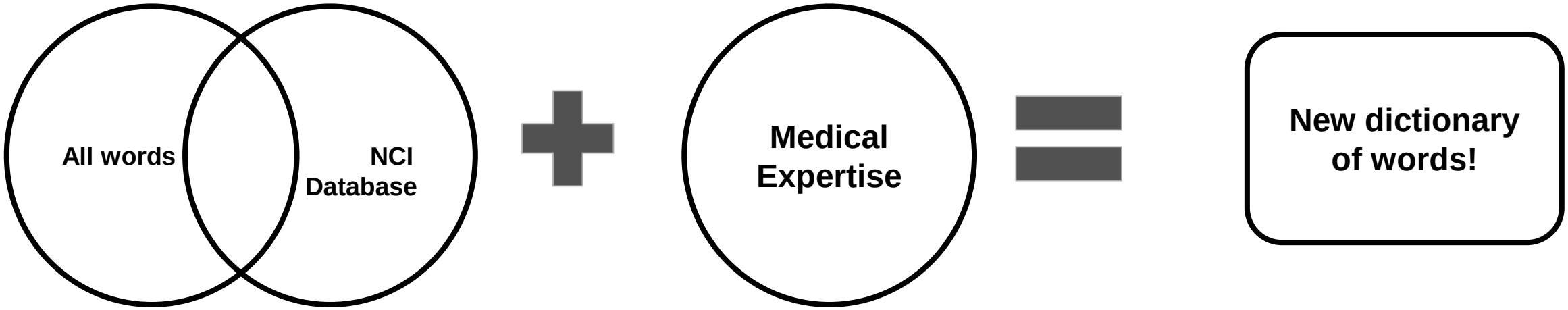
Formatted Generated Sample Pathology Report (from Orchard Pathology Laboratories)

- **Patient Name:**
  Patient, John. | M | DOB: 4/12/1979 | Patient ID :54321-6 | Phone: (123) 555-1234 | EMR: (123) 555-1234
- **Physician Information**:
  James Provider, MD | ABC Medical 400 Royal Drive Anytown USA 12345 | Phone: (123) 555-4321 | Fax: (999) 555-4322
- **Final Diagnosis:**
  A. Ascending Colon: SESSILE SERRATED ADENOMA (POLYP) WITH LOW-GRADE ADENOMATOUS DYSPLASIA
  B. Sigmoid Colon: TUBULAR ADENOMA
- **Comment:**
  Patients with sessile serrated adenomas, especially with cytologic dysplasia, are at increased risk for the development of adenocarcinoma showing microsatellite instability. This progression may occur at a more rapid rate than with traditional adenomas. Complete endoscopic excision is recommended if clinically appropriate. If unresectable, repeat colonoscopy at a shortened interval (1 year), with sampling of suspicious areas or surgical resection possibly warranted.
- **Accession Number**: 12XX0002
  Collection Date: 2/15/2013
  Received Date: 2/15/2013
  Report Date: 2/17/2013 TAT: [26 hours]
- **Specimen:** 2 cm polyp ascending colon 2 mm polyp in sigmoid colon
- **Clinical History**: Screening colonoscopy. Maternal hx of adenocarcinoma of colon age 57
- **Gross Examination**
  A. The first container is labeled “ascending colon.” It contains a polypoid piece of tan mucosal tissue measuring 2.0 cm in greatest dimension. The polyp margin is inked, sectioned, and submitted in cassettes Al and A2.
  B. The second container is labeled “sigmoid colon.” It contains one piece of light tan mucosal tissue 0.2 cm in greatest dimension. Entirely submitted in cassette B.
- **Microscopic Examination:**
  Microscopic Examination performed supportive of the Final Diagnosis

2. Approach

Solution:

Make a special word dictionary!



3. Baseline Model- Counter

3. Baseline Model- counter

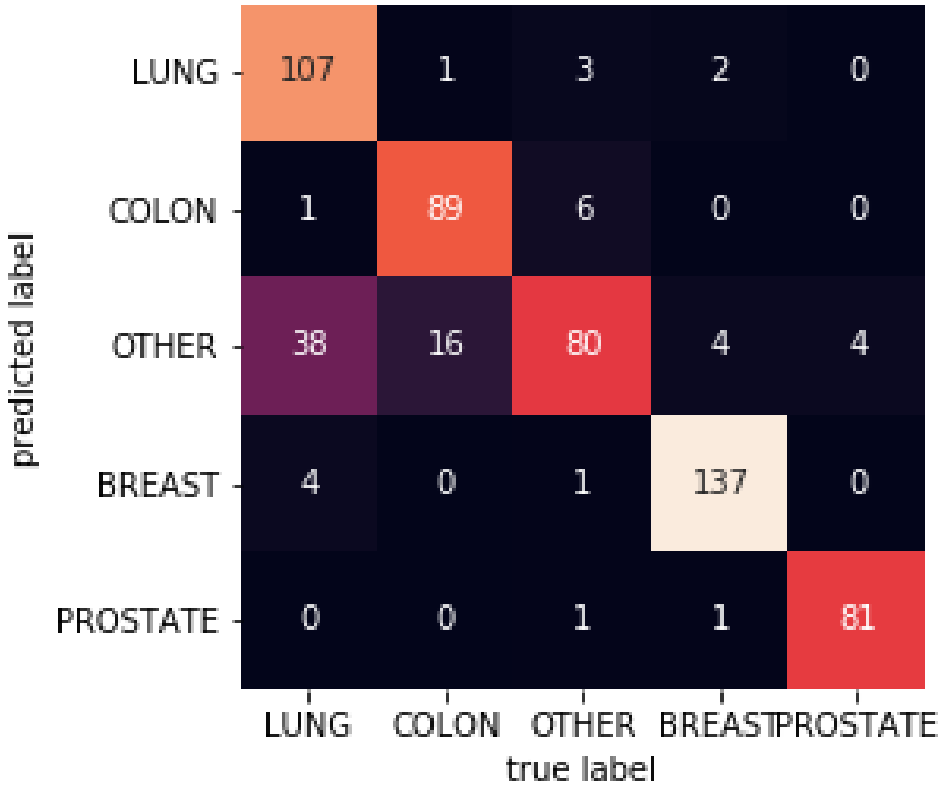
- Count the occurrences of 4 keywords: **Prostate, Lung, Cancer, Breast** in records.
- Classify the site based on the highest occurring keyword

	SITE	Text	Other_count	Colon_count	Lung_count	Breast_count	Prostate_count	Prediction
815	COLON	clinical diagnosis and history bowel obstructi...	0	20	0	0	0	COLON
820	COLON	history: mass in right colon moderately differ...	0	16	0	0	0	COLON
677	LUNG	nan nan nan <clinical hx> lung ca, left. <h...	0	0	16	0	0	LUNG
1032	LUNG	preop diagnosis: right upper lob...	0	0	13	0	0	LUNG
60	BREAST	clinical diagnosis and history the working his...	0	0	0	33	0	BREAST
50	BREAST	clinical diagnosis and history left breast mas...	0	0	0	31	0	BREAST
1134	PROSTATE	clinicalhistory: elevated psa (790.93), prosta...	0	0	0	0	23	PROSTATE
1135	PROSTATE	clinicalhistory: elevated psa. nan finaldiagno...	0	0	0	0	21	PROSTATE
19	OTHER	nan nan nan <clinical info> the patient is a ...	0	0	0	0	0	OTHER
914	OTHER	clinical history the working diagnosis is not ...	0	0	0	0	0	OTHER

3. Baseline Model Results

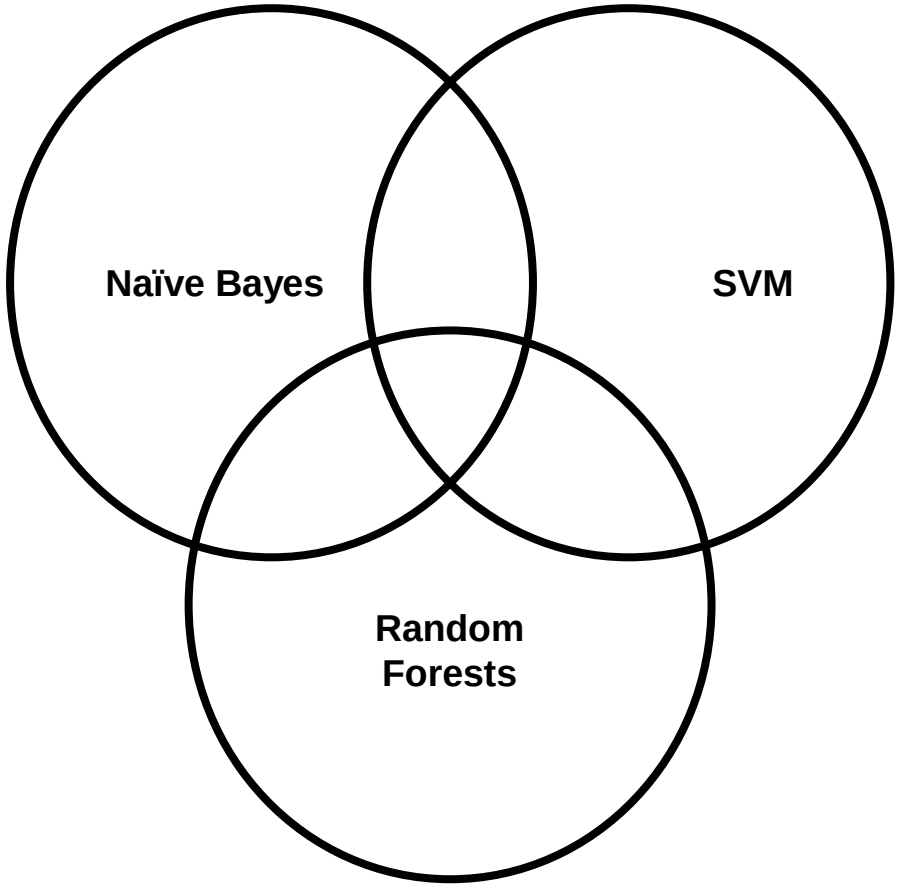
F1 MACRO:
0.86078

MODEL ACCURACY:
85.76%



4. Final Model-
Naive Bayes + SVM + Random Forests

4. Final Model- Naive Bayes + SVM + Random Forests

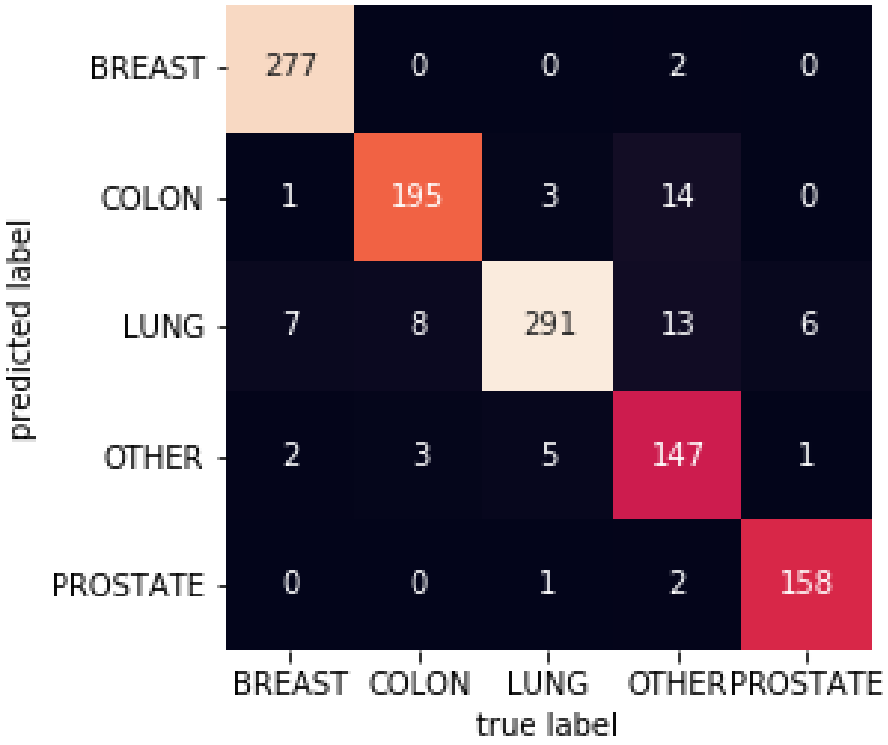


4. Final Model Results

F1 MACRO:
0.936

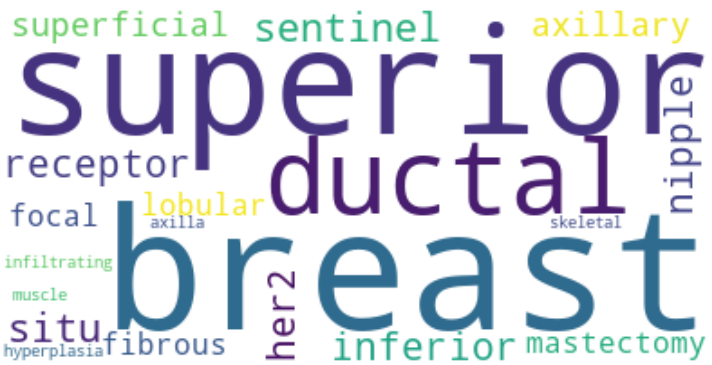
MODEL ACCURACY:
94.09%

Confusion matrix

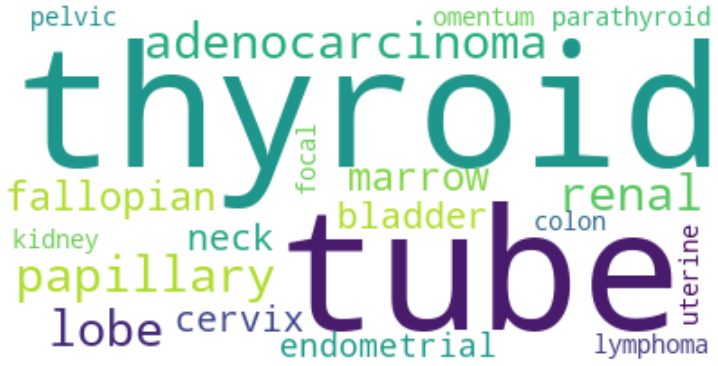


5. Conclusion

PROSTATE



BREAST



OTHER

LUNG

COLON

6. Future work

Challenges of natural language processing	
Challenge	**Example**
Negation	“No evidence of malignancy” in support of an OTHER classification.
Ambiguity with respect to subject	A pathological observation may refer to a historical sample. A LUNG cancer case has the phrase “cancer of the colon” because “the patient has a history of”. One pathology report may describe more than one sample. “No evidence of malignancy” occurs in a report of a cancer case because it refers to a sample from the tumor margin.
Statistical sample size	The ‘OTHER’ class is a union of very different classes. The OTHER class comprises small numbers of samples representing non-cancer as well as cancer of the blood, skin, stomach, etc.
Latent cross classification	
Stochastic independence in the sample	The identity of the registry may be associated with both SITE and usage (confounding).