Medical Transcript Classification

Evan Staffen

Agenda

Business Understanding

Dataset

Modeling

Conclusions

Business Understanding

- Streamline Electronic Medical Records
- More time spent with patients
- Enhance research capabilities
- Improve patient outcomes

Dataset

- Kaggle

  The dataset was obtained from Kaggle and contained 3,714 transcripts from doctor's visits

- Nine different specialties
  - Surgery - 1088
  - General Medicine - 775
  - Cardiovascular - 371
  - Orthopedics - 355
  - Neurology/Neurosurgery - 317
  - Radiology - 273
  - Gastroenterology - 224
  - Urology - 156
  - Gynecology - 155

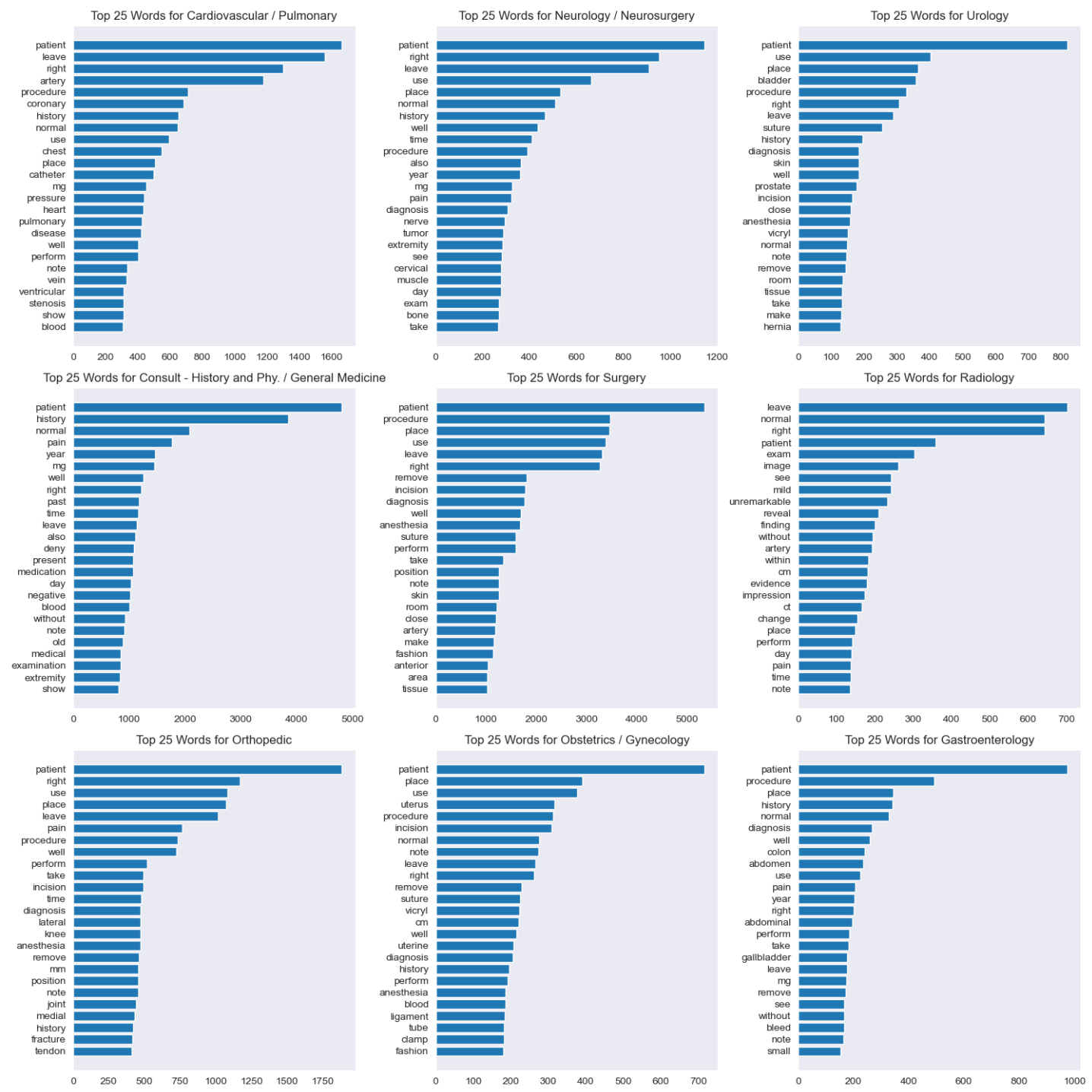
Sample Transcripts

Radiology

'CARDIOLITE TREADMILL EXERCISE STRESS TEST,CLINICAL DATA:, This is a 72-year-old female with history of diabetes mellitus, hypertension, and right bundle branch block.,PROCEDURE:, The patient was exercised on the treadmill to maximum tolerance achieving after 5 minutes a peak heart rate of 137 beats per minute with a workload of 2.3 METS. There was a normal blood pressure response. The patient did not complain of any symptoms during the test and other than the right bundle branch block that was present at rest, no other significant electrographic abnormalities were observed.,Myocardial perfusion imaging was performed at rest following the injection of 10 mCi Tc-99 Cardiolite. At peak pharmacological effect, the patient was injected with 30 mCi Tc-99 Cardiolite.,Gating poststress tomographic imaging was performed 30 minutes after the stress.,FINDINGS:,1. The overall quality of the study is fair.,2. The left ventricular cavity appears to be normal on the rest and stress studies.,3. SPECT images demonstrate fairly homogeneous tracer distribution throughout the myocardium with no overt evidences of fixed and/or reperfusion defect.,4. The left ventricular ejection fraction was normal and estimated to be 78%.,IMPRESSION: , Myocardial perfusion imaging is normal. Result of this test suggests low probability for significant coronary artery disease.'

Cardiology

'2-D M-MODE: , ,1. Left atrial enlargement with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left ventricular ejection fraction of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology of aortic valve, mitral valve, tricuspid valve, and pulmonary valve.,7. PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral and tricuspid regurgitation.,2. Trace aortic and pulmonary regurgitation.'



Word Frequency by Specialty

- Removing common words negatively impacted model results
  - e.g. patient, place, use, procedure, etc.
- Clear differences amongst specialties
- General Medicine and Surgery
  - Common words not highly distinguishable for classification

Quick Guide

Modeling Process

NLTK

Tokenize documents

Lemmatize

Vectorize

Modeling

Gensim + SpaCy

Tokenize documents

Lemmatize

Vectorize text

NMF Model

- Create topic weights per document

Modeling

Word2Vec

Clean with Gensim + SpaCy

Train Custom Word2Vec Model

Convert to Vectors

- 100 dimensional

LSTM Modeling

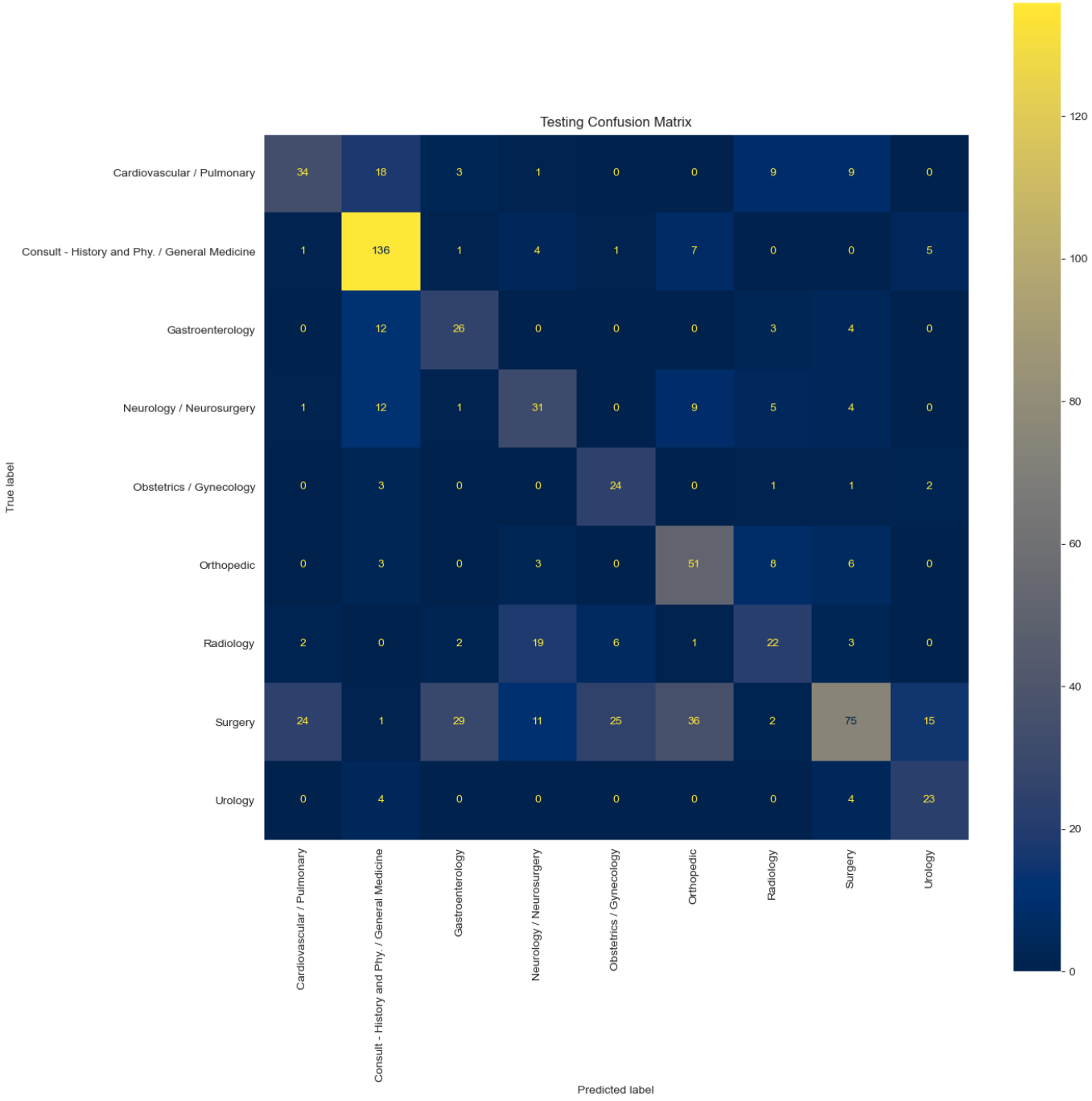
GloVe

Clean with Gensim + SpaCy

Convert with pre-trained Vectors

- Used wiki-gigaword-100

LSTM Modeling



Best Model

- Gensim + SpaCy
  - TfidfVectorizer
  - NMF model Topic Weights
- Logistic Regression

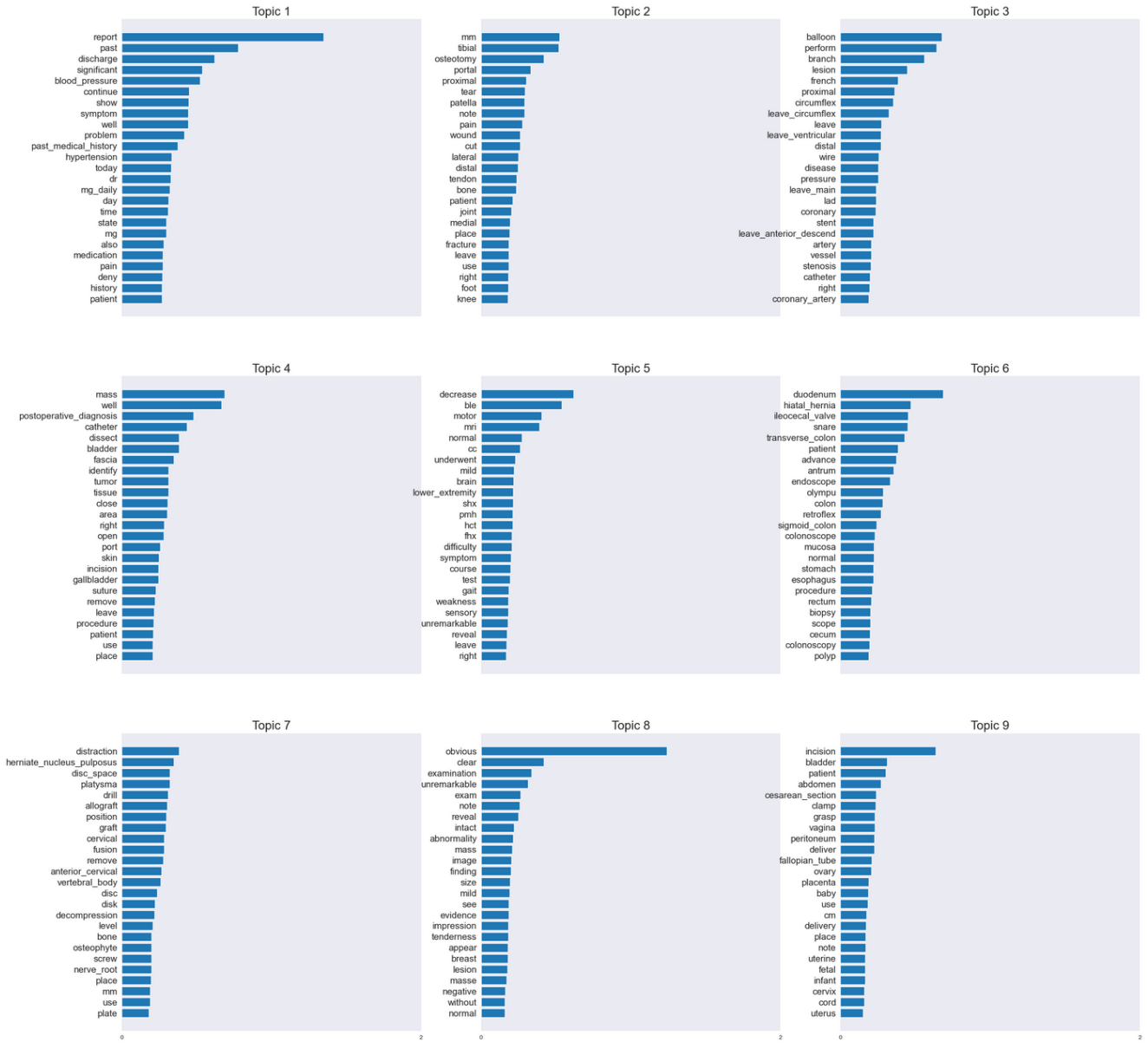
Metrics

- **Accuracy: 56.80%**
- **Precision: 59.45%**
- **Recall: 56.80%**
- **F1-Score: 55.45%**

NMF Model Topic-Word-Weights

- Some clear topics
  - Topic 2 likely Orthopedic
  - Topic 3 likely Cardiovascular
  - Topic 5 likely Neurology
  - Topic 6 likely Gastroenterology
  - Topic 8 likely Radiology
  - Topic 9 likely Urology

- Some unclear topics
  - Topic 1 possibly General
  - Topic 4 possibly Surgery
  - Topic 7 possibly Gynecology



Conclusions

Improvements needed

Differentiates specialties well, surgery and general medicine categories caused errors

For practical implementation, scores need to be improved

Acquire more data

More data could lead to higher scores and viable LSTM models

Help differentiate between broad categories

Pre-trained vectors

More research into deep learning and word2vec modeling

GloVe vectors missing medical terminology

Questions?

Contact

https://github.com/evanstaffen/Medical-Transcript-Classification

https://www.linkedin.com/in/evan-staffen-a74045207/

evan.staffen@gmail.com