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

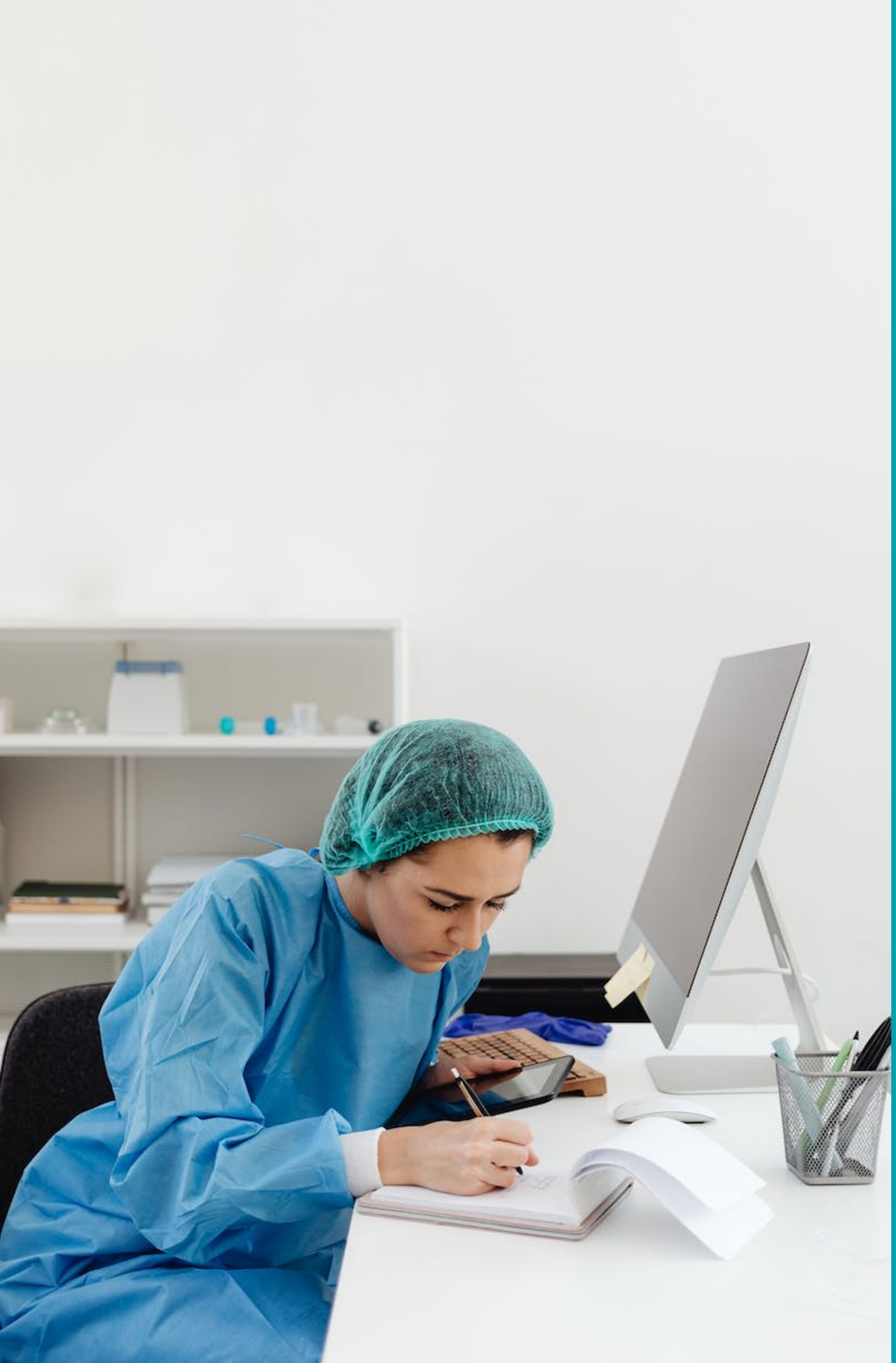
Improve patient outcomes

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



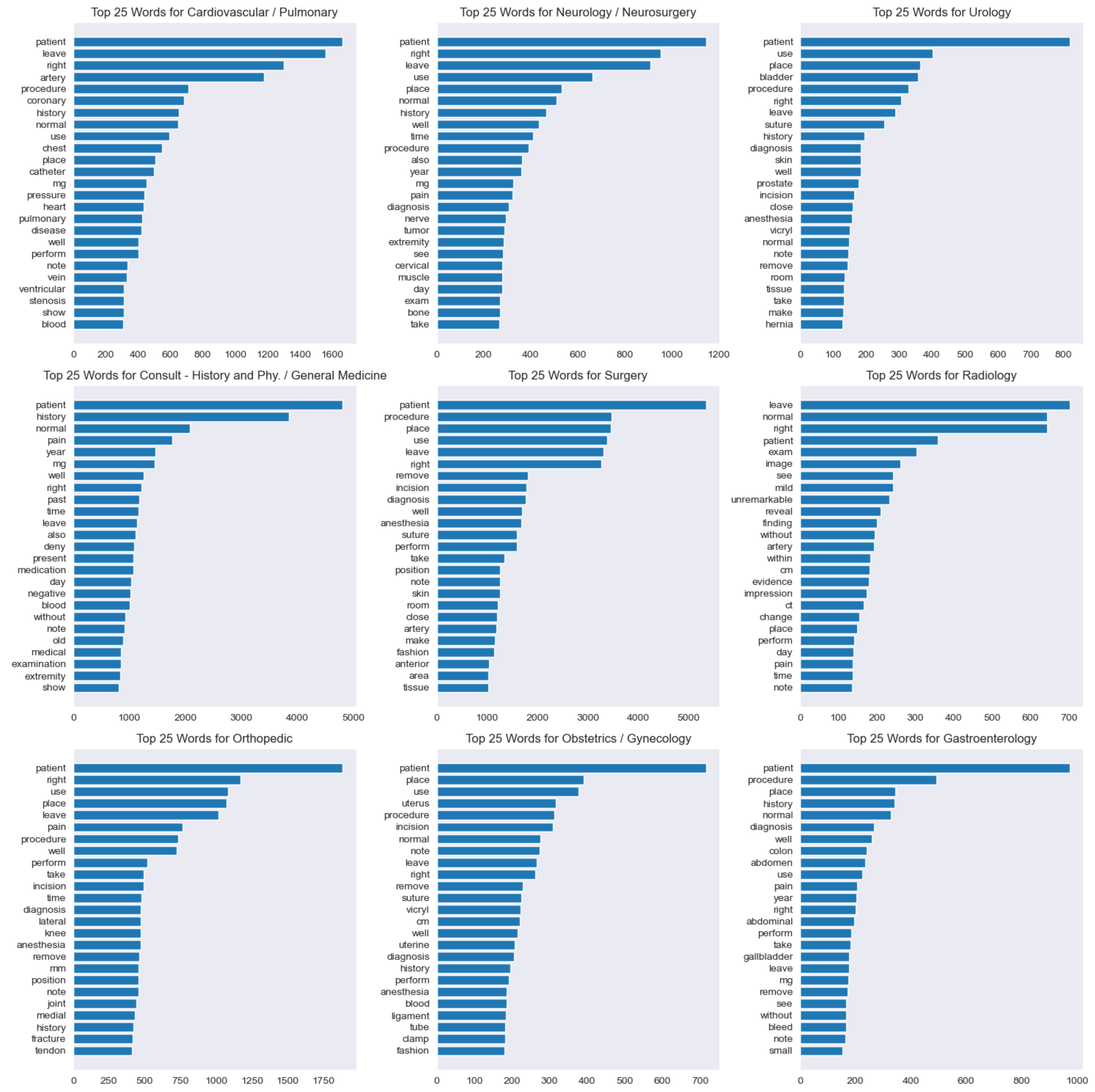
Sample Transcripts

Cardiovascular

'2-D M-MODE: , ,1. Left atrial enla ial diameter of 4.7 cm.,2. Normal size right and left ventricle., 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.'

Urology

"PREOPERATIVE DIAGNOS al hernia.,POSTOPERATIVE DIAGNOSIS: , Umbilical hern URE PERFORMED: , Repair of umbilical hernia.,ANESTHESIA: , General.,COMPLICATIONS: , None.,ESTIMATED BLOOD LOSS: , Minimal.,PROCEDURE IN DETAIL: ,The patient was prepped and draped in the sterile fashion. An infraumbilical incision was formed and taken down to the fascia. The umbilical hernia carefully reduced back into the cavity, and the fascia was closed with interrupted vertical mattress sutures to approximate the fascia, and then the wounds were infiltrated with 0.25% Marcaine. The skin was reattached to the fascia with 2-0 Vicryls. The skin was approximated with 2-0 Vicryl subcutaneous and then 4-0 Monocryl subcuticular stitches, dressed with Steri-Strips and 4 x 4's. Patient was extubated and taken to the recovery area in stable condition."



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

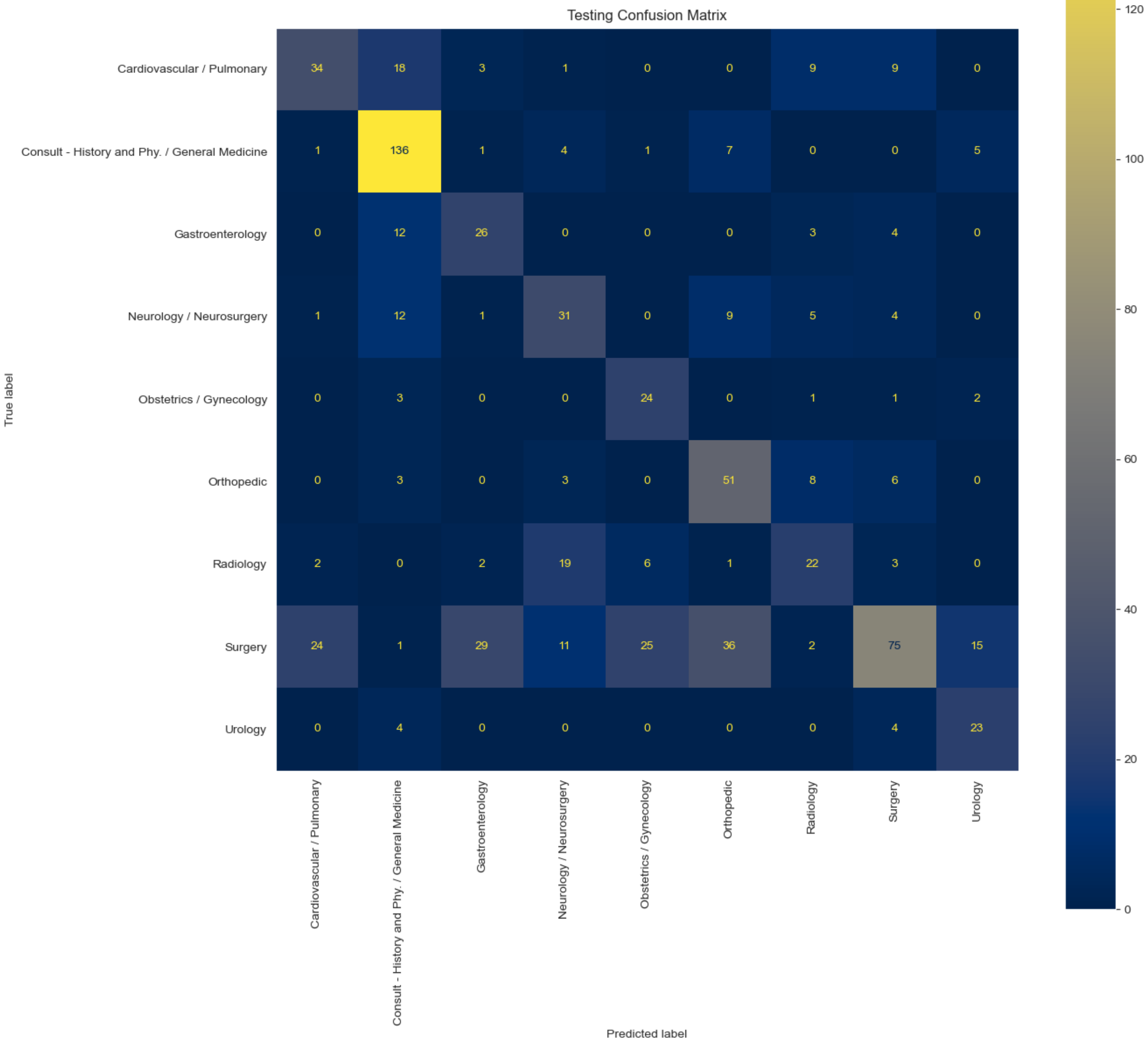
GloVe

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

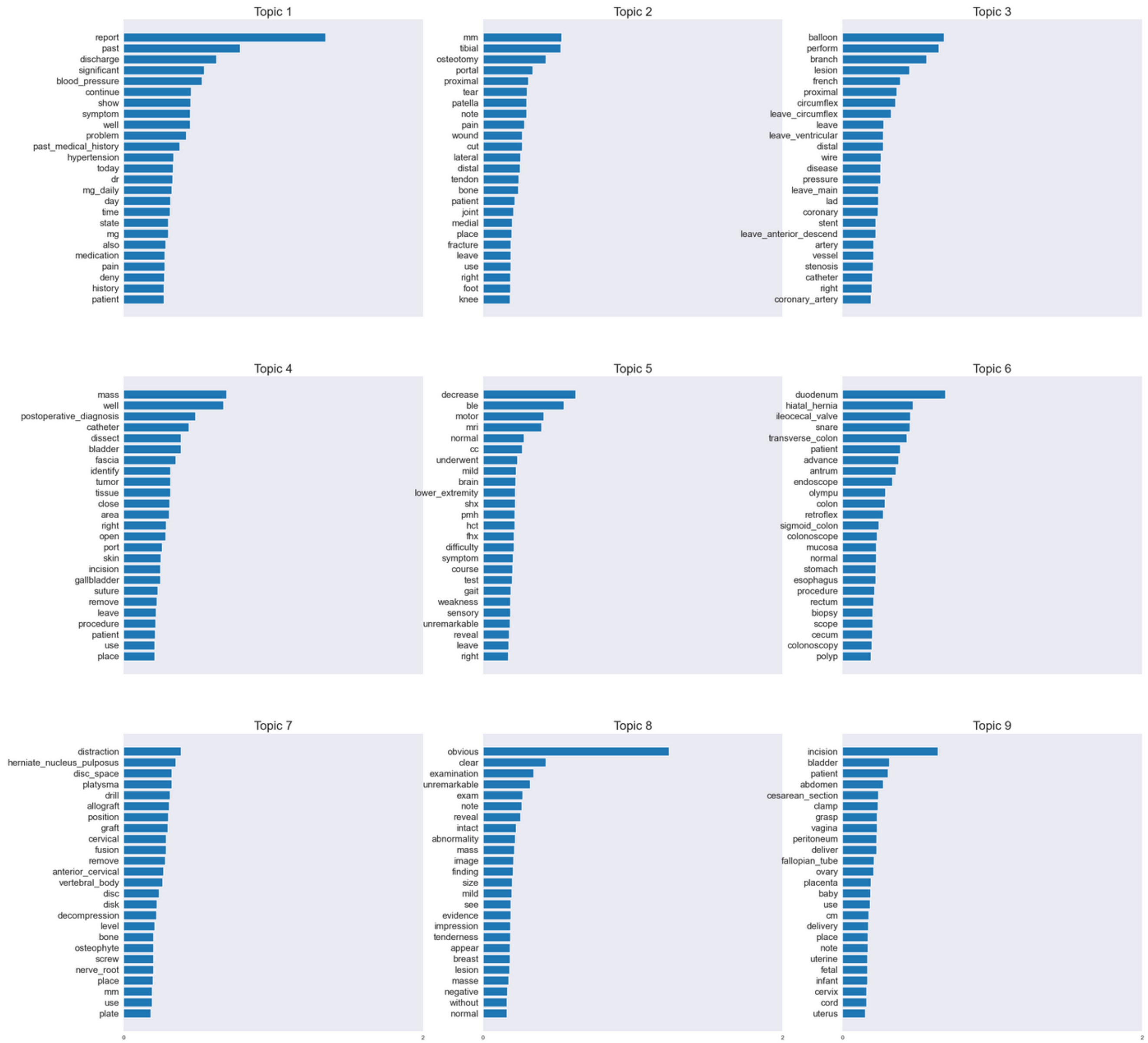
Metrics

- **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