CLASSIFYING MEDICAL DISCHARGE RECORDS WITH SUBJECTIVE FEATURES

Nick Waltner

Mike Roylance

- Healthcare in the United States is in a major state of transformation.
- For the first time in the nation’s history, it has been mandated that every individual be covered under health insurance. With this change, hospitals are being accountable for even more efficient and effective services.
- Healthcare data is complex. Typically medical professionals store information about a patient in free text fields.
- It is estimated that up to 70% of important information about a patient is stored here.

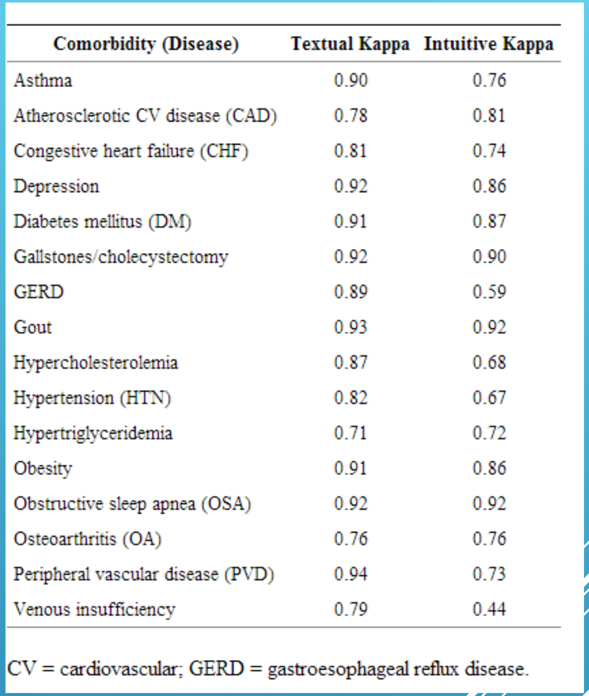
WHY IS THIS RELEVANT?

- We worked on a medical NLP data set from the 2008 Obesity i2b2 competition.
- This data set consists of 1237 patient discharge records written by medical professionals describing the health history and procedures done to the patient during their stay.
- Partners HealthCare in Boston annually holds NLP competitions.

WHAT DID WE WORK WITH?

WHAT DID WE WORK WITH?

- The 2008 Obesity Challenge data set consists of 1237 patient discharge records written by medical professionals describing the health history and procedures done to the patient during their stay.
- These records have been classified with certain diseases, or co-morbidities.
    - A patient diagnosed with depression could also be diagnosed with gout and/or diabetes.
    - A disease is classified as Present, Absent, Unmentioned or Questionable.
    - There are two types of annotations as well, textual or intuitive
- Each record has been annotated by medical professionals with an average 72% kappa agreement.



Comorbidity (Disease)	Textual Kappa	Intuitive Kappa
Asthma	0.90	0.76
Atherosclerotic CV disease (CAD)	0.78	0.81
Congestive heart failure (CHF)	0.81	0.74
Depression	0.92	0.86
Diabetes mellitus (DM)	0.91	0.87
Gallstones/cholecystectomy	0.92	0.90
GERD	0.89	0.59
Gout	0.93	0.92
Hypercholesterolemia	0.87	0.68
Hypertension (HTN)	0.82	0.67
Hypertriglyceridemia	0.71	0.72
Obesity	0.91	0.86
Obstructive sleep apnea (OSA)	0.92	0.92
Osteoarthritis (OA)	0.76	0.76
Peripheral vascular disease (PVD)	0.94	0.73
Venous insufficiency	0.79	0.44

CV = cardiovascular; GERD = gastroesophageal reflux disease.

- To explore the use of subjective language in medical text with the goal of better identifying hidden patterns of sentiment that correlate to the diagnosis.
- To our knowledge, sentiment analysis in the medical domain hasn't been explored.
- This data is produced by humans, even with the goal of being objective, subjective descriptions might exist.
  - Another topic, identifying sarcasm in the medical domian.

WHY DID WE CHOOSE TO WORK ON THIS?

- Downloaded the data from i2b2.org. (XML documents)
- Built a data warehouse for storage and analysis
  - Storage and “slicing and dicing” the data was done in MySQL
  - ETL was done in Python, Perl, and R
  - Stanford NLP Parser
  - Classification tools were
    - Python SKLearn (Random Forest, MaxEnt)
    - Mallet (MaxEnt)
  - Feature selection and modification was done in Python.

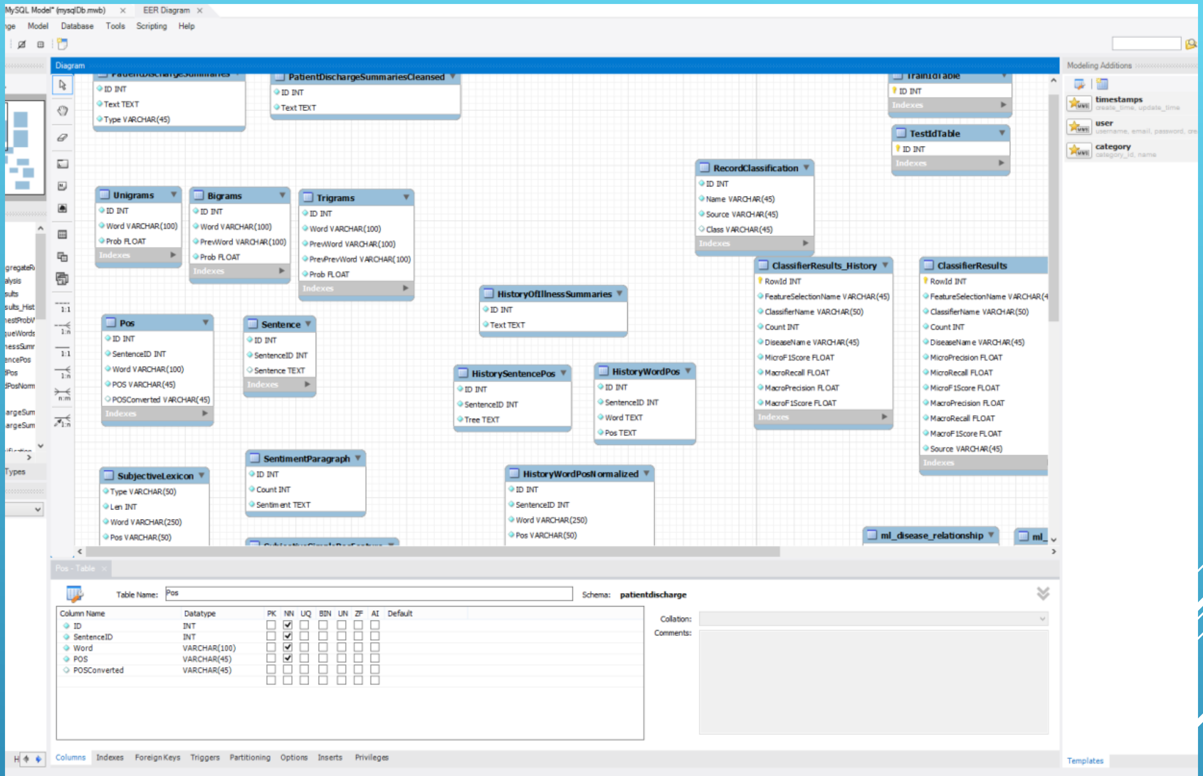
HOW WE DID IT

- Extracted the text documents (with their IDs) from the XML files into the database
- Converted the text documents into POS tagged sentences.
  - Extracted the entire instance for evaluation.
  - Extracted small subsections of text that contained fewer medical ontology words.
- Uploaded Subjective Lexicon:
  - http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- Loaded data from the database into python for transformation

PYTHON / PYTHON / R

MYSQL

- O(log(n)) searching
- Open source, hardened community
- SQL is the lingua franca of data
- Hosted from my house



Pos - Table

Table Name: Pos

Schema: **patientdischarge**

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	Default
ID	INT	□	☑	□	□	□	□	□	
SentenceID	INT	□	☑	□	□	□	□	□	
Word	VARCHAR(100)	□	☑	□	□	□	□	□	
POS	VARCHAR(45)	□	☑	□	□	□	□	□	
POSConverted	VARCHAR(45)	□	□	□	□	□	□	□	
		□	□	□	□	□	□	□	

Collation:

Comments:

Columns Indexes Foreign Keys Triggers Partitioning Options Inserts Privileges

- With the data in the database, separated out the training and the test instances.
- Trained instances with several features
  - Custom Word Count feature.
  - Subjective Sentiment Simple
    - Positive / Negative based on word / pos
  - Subjective Sentiment Complex
    - Positive / Negative with strong/weak polarity on word / pos
  - Subjective Sentiment Word
  - Subjective Sentiment Word Polarity
  - Unique Words
  - Highest Probability Words

MACHINE LEARNING

- Will compile by tonight

RESULTS