# Classifying Physician Needed Response Questions on Online Health Forums

## (Final Project of SI699: Big Data Analytics)

Kristen McGarry[1,2], Deahan Yu[1,2]

[1] School of Information, [2] School of Public Health, University of Michigan

## Introduction

Problem: Patients seek medical advice on online platforms, but not all questions need a response from a medical professional. With rising rates of physician fatigue and the demand for physician time, it is difficult for physicians to respond to questions on these online platforms.

Solution: We created a form of 'weak AI', an automatic filter, to help sort through questions that need a doctors response.

*"I have two nodules one on each side of my thyroids. my labs came back normal but the test for thyroid peroxidase ab came back <10.0 and standard range is 0-35 is this ok."*

*"What exercise do u plan on doing next week?"*

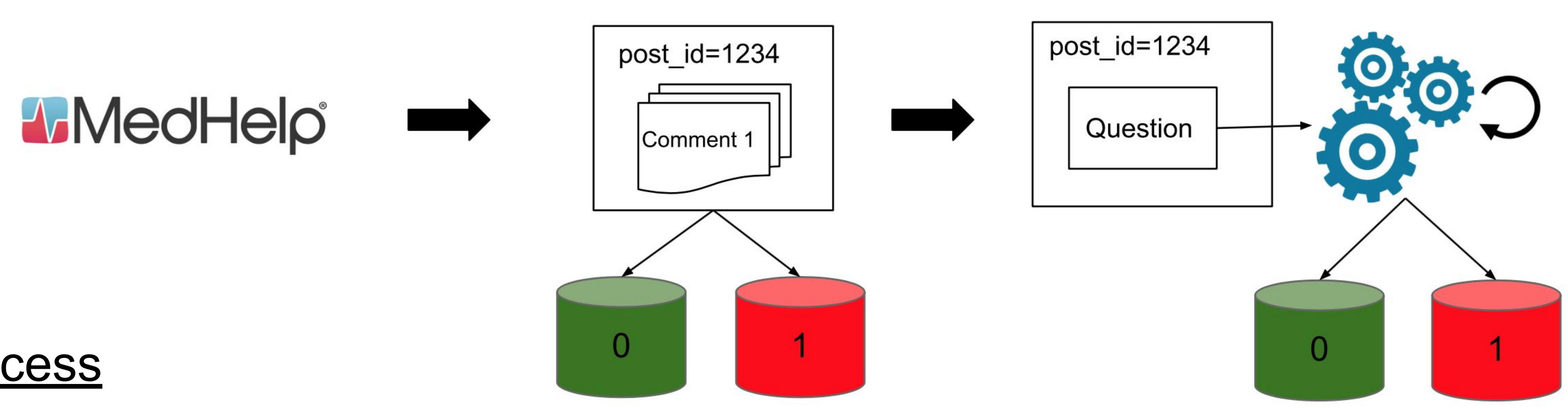
*"I had a root canal done over 2 years ago ... i get sinus pressure on the left area of my face especially above the molar with the root canal, ... and i even have swullon cheek that i can notice, which appears unidentified by others, but i feel they say this 2 make me feel better? any suggestions?"*

## Process

**Data:**
1,403,220 questions
2,366,387 follow-up comments

**Task 1:**
Question Labeling Task based on Comments

**Task 2:**
Question Classification Task




## Task 1. Question Labeling Task

| | | | |
|---|---|---|---|
| Initial Regex | Random 200 comments | | |
| Labeled 1's and 0's | Cohen's Kappa = 0.73 | | |
| Finalized Regex | Patterns & Errors | | |
| | precision | recall | f1-score |
| ~23% labeled as 1's | 0.76 | 0.91 | 0.83 |

## Task 2. Question Classification

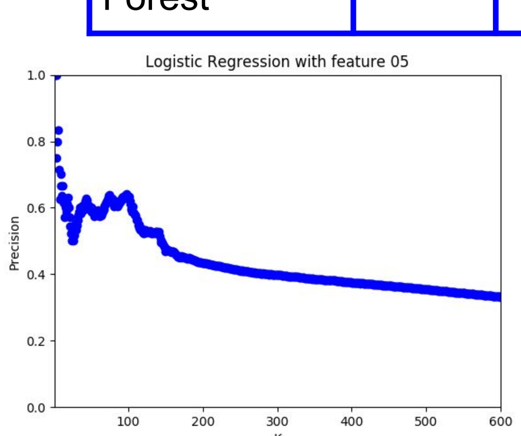
Pre-text processing: stopword removal & "num" for numbers

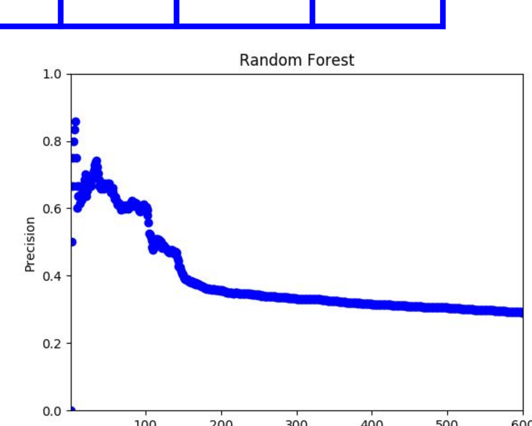
### Standard Models

For our initial approach, we used a supervised learning technique with bag of words & TF-IDF (optional). Below are the top 2 performing models.

Feature Selection:
Start: 571,045 features
Chi-Squared, p-value <0.05: 29,821 features

| Top Models | acc | roc/auc | prec | recall | f1 |
|---|---|---|---|---|---|
| Logistic Regression | 0.75 | 0.66 | 0.68 | 0.75 | 0.66 |
| Random Forest | 0.74 | 0.60 | 0.66 | 0.74 | 0.67 |



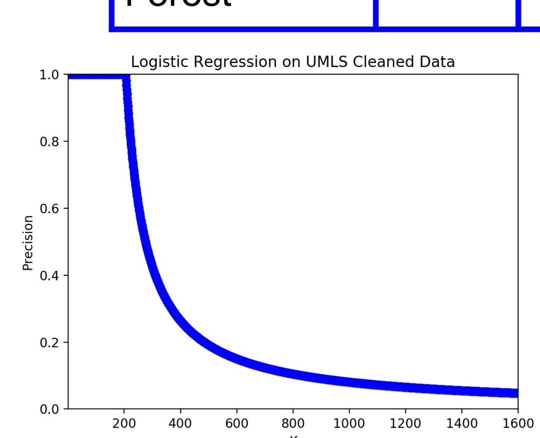

Top 20 Features:
back, chest, condom, doctor, feel, feeling, go, heart, hiv, like, normal, pain, risk, sex, side, sometimes, started, stomach, symptoms, worse

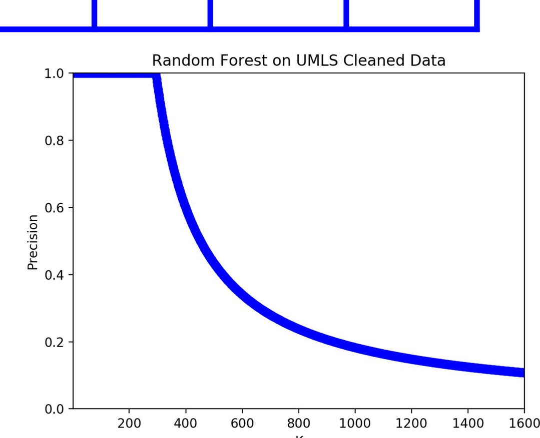
### UMLS Models

For this approach, we utilized Quick UMLS to match medical concepts to the tokenized questions. We used a supervised learning technique with bag of words & TF-IDF. Below are the top 2 performing models.

Feature Selection:
Start: 29,134 features
Chi-Squared, p-value <0.05: 7,801 features

| Top Models | acc | roc/auc | prec | recall | f1 |
|---|---|---|---|---|---|
| Logistic Regression | 0.75 | 0.66 | 0.68 | 0.75 | 0.66 |
| Random Forest | 0.74 | 0.60 | 0.66 | 0.74 | 0.67 |




Top 20 Features from UMLS:
abdominal, breathing, chest, condom, condoms, constipation, depression, discharge, dizziness, dizzy, dog, exposure, feeling, heart, medication, medications, pain, pressure, sex, stomach

## Conclusions

1. UMLS models performs better than standards models
   a. Incorporating UMLS does help
2. Random Forest with UMLS performs the best
3. But, precision still has room for improvement

## Lessons

1. "Occam's Razor"
   a. Sometimes the simplest model works the best
2. Labeling limitations and biases
   a. Need physician's input
3. When working with big data, runtime is the bottleneck

## Future Work

1. Partner with researchers to continue the project forward!
2. Continue training models to achieve higher support
3. Implement the neural network models (LSTM, CNN)
4. Develop "better" RegEx for question labeling task
   a. Incorporate physician's input

This is a final project presentation of SI 699: Big Data Analytics. We thank our advisor Qiaozhu Mei, and the provider of our scraped dataset, Danny Wu.

**Interested in helping? Contact: mcgarryk@umich.edu or deahanyu@umich.edu**