Identifying Patient Smoking Status from Medical Discharge Records

Project Proposal for CS7180 – Special Topics in Artificial Intelligence Summer '19
submitted by Akriti Chadda

Proposed Dataset

Available at : https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

The data consists of discharge summaries from Partners HealthCare and preprocessed at the i2b2 Center (Informatics for Integrating Biology and the Bedside). The records are in XML format and de-identified to protect patient security and privacy. Institutional review boards of Partners HealthCare, Massachusetts Institute of Technology, and the State University of New York at Albany approved the challenge and the data preparation process.

Why is this important?

Electronic Medical Records / Electronic Health Records provide valuable information regarding the diagnoses, treatment and response history for patients. However, much of the information available in these documents is fragmented and unstructured in nature, and doesn't always follow the rules of grammar. These records are also embedded with the tobacco use history of the patients, something that is the root cause of many pulmonary and coronary diseases such as myocardial infarction as well as a variety of cancers. Given the nature of how these medical records have been filled in the past, clinicians usually have to spend hours in going through these documents manually to determine the tobacco use of patients. This is very expensive monetarily and in terms of the time spent in doing it. Hence, automatic identification of tobacco use becomes a challenge as well as an opportunity for the medical NLP community.

Big Question

Can a patient's smoking status be automatically and correctly identified from their clinical records?

Sample

```
977146916
HLGMC
2878891
022690
01/27/1997 12:00:00 AM
CARCINOMA OF THE COLON .
Unsigned
DIS
```

Report Status :

Unsigned

Please do not go above this box important format codes are contained .

DISCHARGE SUMMARY

ARF32 FA

DISCHARGE SUMMARY NAME :

GIRRESNET , DIEDREO A

UNIT NUMBER :

075-71-01

ADMISSION DATE :

01/27/1997

DISCHARGE DATE :

01/31/1997

PRINCIPAL DIAGNOSIS :

Carcinoma of the colon .

ASSOCIATED DIAGNOSIS :

Urinary tract infection , and cirrhosis of the liver .

HISTORY OF PRESENT ILLNESS :

The patient is an 80-year-old male , who had a history of colon cancer in the past , resected approximately ten years prior to admission , history of heavy alcohol use , who presented with a two week history of poor PO intake , weight loss , and was noted to have acute on chronic Hepatitis by chemistries and question of pyelonephritis .

He lived alone but was driven to the hospital by his son because of reported worsening and general care and deconditioning .

Emergency Department course ; he was evaluated in the emergency room , found to be severely cachectic and jaundiced .

He was given a liter of normal saline , along with thiamine , folate .

An abdominal ultrasound was performed showing no stones .

Chest x-ray revealed clear lungs and then he was admitted to Team C for management .

PAST MEDICAL HISTORY :

Cancer , ten years prior to admission , status post resection .

MEDICATIONS ON ADMISSION :

Folic acid .

ALLERGIES :

None .

FAMILY HISTORY :

Not obtained .

SOCIAL HISTORY :

Lives in Merca .

Drinks ginger brandy to excess , pipe and cigar smoker for many years .

PHYSICAL EXAMINATION :

In general was a cachectic , jaundiced man .
bloodpressure :
124/60 , 97.4 , 84 , 22 for vital signs .
head , eyes , ears , nose and throat :
notable for abscess ulcers on the lower gums .
He was edentulous .
Neck was supple , lungs were clear except for some scattered mild crackles .
Cardiac :
tachycardic with a II / VI systolic ejection murmur .
Belly was tender in the right upper quadrant .
Liver edge , thickened abdominal wall was palpable .
No inguinal nodes .
Rectal was guaiac negative .
On mental status exam , he was somnolent but arousable .
Oriented to name , year , and hospital .
Skin was jaundiced .
LABORATORY DATA :
Notable for a BUN and creatinine 14 and 1.8 , phosphorous of .5 , magnesium 1.2 , albumin 2.1 .
elevated liver function tests , bilirubin of 14 direct , 17 total .
uric acid 11.4 , alkaline phosphatase 173 , serum glutamic oxaloacetic transaminase 309 , amylase 388 .
His urinalysis showed 10-20 granular casts and 10-20 white blood cells , 3-5 red blood cells , 5-10 whites , 3-5 white blood cells cast .
The white blood cell was 8.5 , hematocrit 34 .
platelet count 74 .
5% bands on differential .
prothrombin time 14.9 , partial thromboplastin time 35 .
HOSPITAL COURSE AND TREATMENT :
The patient was admitted to the Staviewordna University Of Medical Center .
His mental status proceeded to decline as he became more sleepy and less arousable and confused .
His Hepatitis worsened , liver failure progressed with his coagulopathy worsening .
His renal status also decreased with a drop in urine output , became more shortness of breath as he developed some pulmonary edema .
A head computerized tomography scan was planned to evaluate his change in mental status , but after an extensive discussion with the son , who felt that he and other family members wanted to maximize the patient 's comforts and avoid heroic measures in the event of further deterioration , plans were made to make the patient as comfortable as possible .
He was continued on antibiotics , and oxygen , and morphine , and small amounts of Dopamine , and at 4 AM on January 31 , was pronounced dead .
_____ AJO C. CUCHKOTE , M.D.
TR :

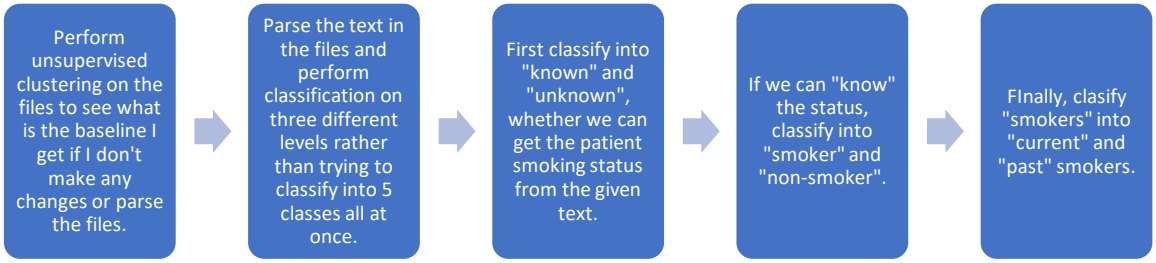
```
tfv
DD :
09/08/1997
TD :
10/13/1997 3:47
Pcc :
AZEL USANNE WALL , M.D.
[ report_end ]
```

Having to manually go through this record would take some time, and imagine having to go through tens of them everyday. Whereas, automated classifying of this record into a 'SMOKER' category can save so much time that can be spent in actually helping the patient. This was a simple example, in that it had the words "cigar smoker" but the description can be as direct as that, or as abstract as "past tobacco use", "occasional fagging" etc. The goal is to classify patients into one of the following 5 categories: past smoker (P), current smoker (C), smoker (S), non-smoker (N), and unknown (U).

Why am I excited about this?

1) I am very passionate about the interface of computer science and healthcare.
2) I want to leverage my past experience in biomedical engineering and neuroscience, and use the new skills I am gaining in Machine Learning to do cool things in this space.
3) I think it's a cool problem to find a solution to
4) IF, and it's a BIG IF, I am able to use the techniques we have learned in class efficiently to classify the data, I might try to build a deep learning model and see if the accuracy increases (then it helps me get my feet wet in deep learning too)\

Proposed Plan



Of course, all of this will be followed by the evaluation of my model against the ground truth that has been provided as a part of my data-set.

User-Facing service

While I am not a 100% sure yet about the implementation, I am thinking on the lines of having a textbox where users can paste the discharge summary text (or if I am able to figure it out, upload a file of the discharge summary) and getting an output whether the patient falls in one of the 5 above-mentioned categories. I think of it as an add-on to Electronic Health Record services, one of them being patient classification.