Combining NLP and Tabular Features

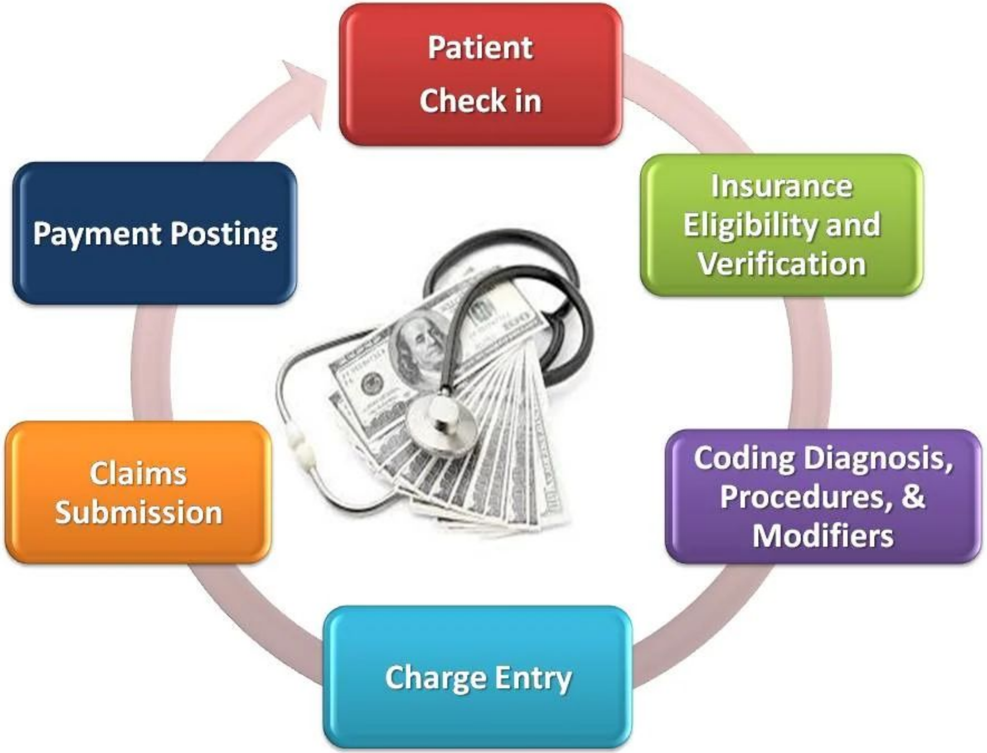
Mapping Clinical Entities to the relevant Section Headers

Sagar Dawda
Data Scientist III
Episource India Pvt Ltd

About Me

- Data Scientist III at Episource India Pvt Ltd
- NLP at large scale on Clinical Text
- Some recent projects:
    - Clinical entities mapping to relevant section headers
    - Breaking clinical documents into appropriate report boundaries
    - Domain specific ontology search for diseases

The cycle of Medical Coding


Patient Check in
Insurance Eligibility and Verification
Coding Diagnosis, Procedures, & Modifiers
Charge Entry
Claims Submission
Payment Posting

Refer to the link for more info - https://www.m-scribe.com/blog/6-signs-you-have-a-great-medical-billing-process

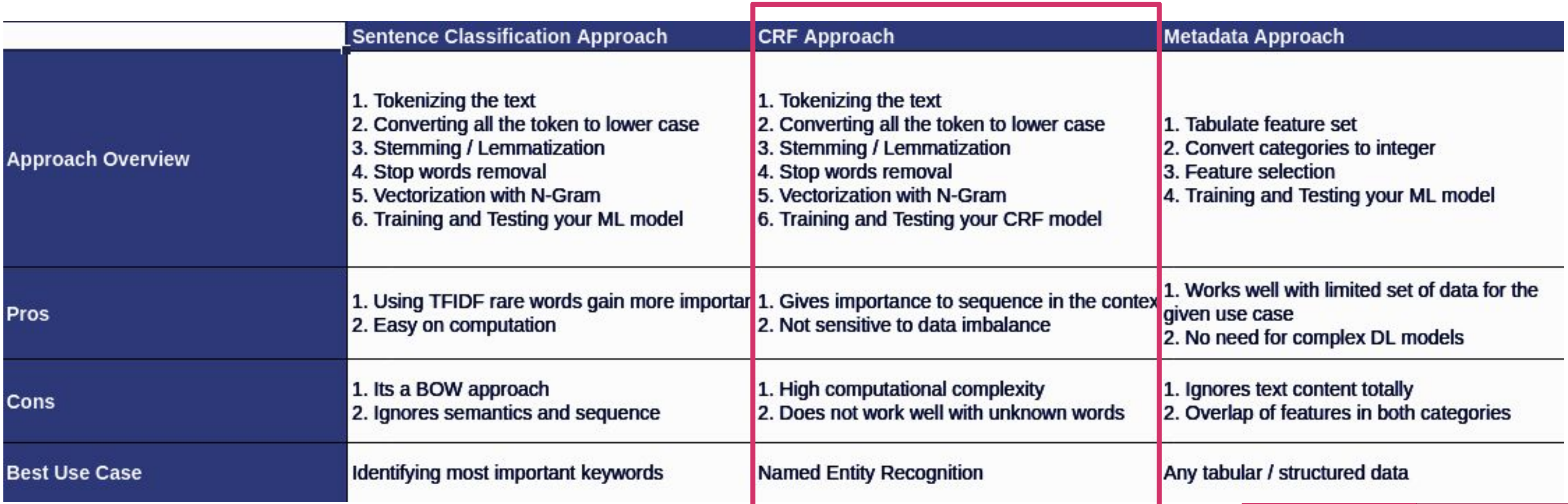
Objective of the project

- ➔ Section Headers play a critical role in determining the relevance of the Disease
- ➔ Incorrect Disease identification from an irrelevant section can lead to revenue loss of the Insurers
- ➔ Identification alone is not helpful for the project
- ➔ Relevant mapping is pivotal for retention / elimination of Disease for claims
- ➔ Project was built to identify the accurate section headers and map relevant entities with them

Impact of Section Headers

- Determine if the section is acceptable for capturing code
- Determine if the condition is current or historic
- Determine if the clinical document follows a SOAP format for capturing information
- Incorrect mappings can lead to the following consequences:
  - Additional payments leading to revenue loss - Type 1 errors
  - Incorrectly rejected claims - Type 2 errors

Initial Approaches



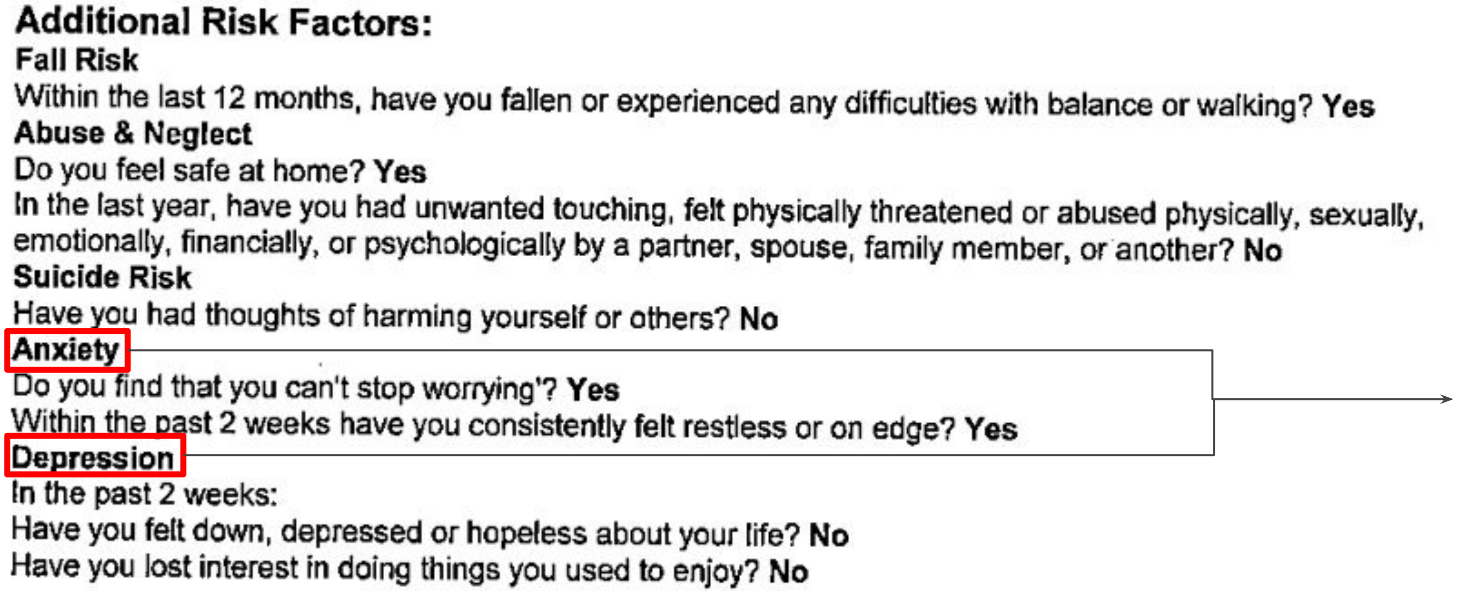
	Sentence Classification Approach	CRF Approach	Metadata Approach
Approach Overview	1. Tokenizing the text 2. Converting all the token to lower case 3. Stemming / Lemmatization 4. Stop words removal 5. Vectorization with N-Gram 6. Training and Testing your ML model	1. Tokenizing the text 2. Converting all the token to lower case 3. Stemming / Lemmatization 4. Stop words removal 5. Vectorization with N-Gram 6. Training and Testing your CRF model	1. Tabulate feature set 2. Convert categories to integer 3. Feature selection 4. Training and Testing your ML model
Pros	1. Using TFIDF rare words gain more importar 2. Easy on computation	1. Gives importance to sequence in the contex 2. Not sensitive to data imbalance	1. Works well with limited set of data for the given use case 2. No need for complex DL models
Cons	1. Its a BOW approach 2. Ignores semantics and sequence	1. High computational complexity 2. Does not work well with unknown words	1. Ignores text content totally 2. Overlap of features in both categories
Best Use Case	Identifying most important keywords	Named Entity Recognition	Any tabular / structured data

Shortlisted Approach

Drawback of the existing approach

- Section Headers are not sequence problems
  - Chief Complaint - Relevant Section Header
  - Patient is taking medications regularly for all the conditions listed in chief complaint section
- Too many Type 1 errors
- Text sequence may not match the PDF format
  - Section headers are usually not a part of any sentence

Examples



Additional Risk Factors:

Fall Risk

Within the last 12 months, have you fallen or experienced any difficulties with balance or walking? **Yes**

Abuse & Neglect

Do you feel safe at home? **Yes**

In the last year, have you had unwanted touching, felt physically threatened or abused physically, sexually, emotionally, financially, or psychologically by a partner, spouse, family member, or another? **No**

Suicide Risk

Have you had thoughts of harming yourself or others? **No**

Anxiety

Do you find that you can't stop worrying'? **Yes**

Within the past 2 weeks have you consistently felt restless or on edge? **Yes**

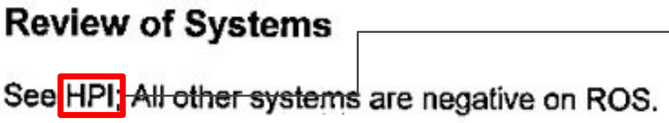
Depression

In the past 2 weeks:

Have you felt down, depressed or hopeless about your life? **No**

Have you lost interest in doing things you used to enjoy? **No**

Section Headers incorrectly classified as Disease



Review of Systems

See HPI; All other systems are negative on ROS.

Incorrect Header

Additional Challenges

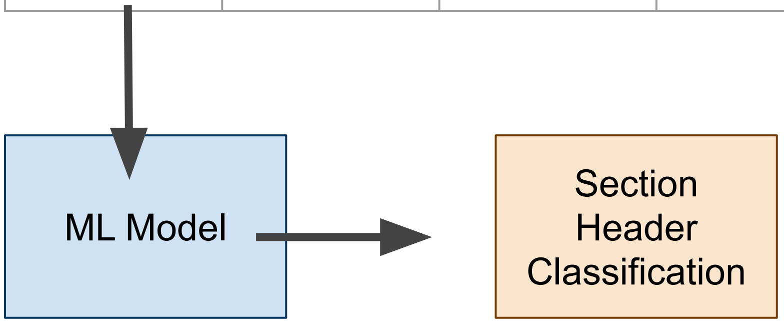
- ➔ Data from a table within PDF chart
- ➔ Table headers
- ➔ OCR errors during PDF to Text conversion
- ➔ Alignment and text sequencing mis-match

The Solution

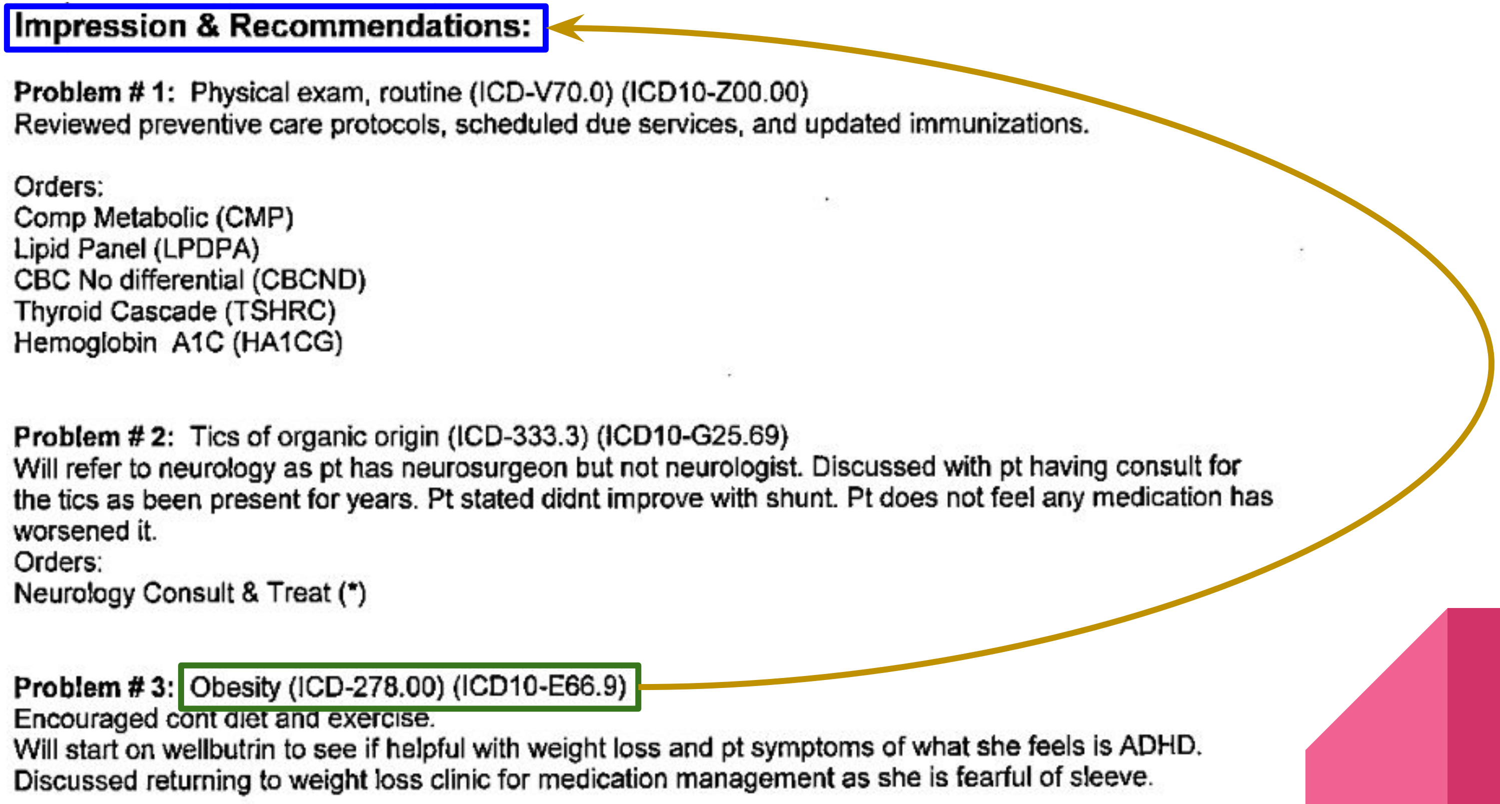
- ➔ Combine Text and Metadata Features:
- ➔ Text features:
  - ◆ Lower case text
  - ◆ Match with known headers
  - ◆ Spell correction
- ➔ Metadata:
  - ◆ Font size
  - ◆ Font style
  - ◆ Position info

Sample of final matrix passed to ML Model

Sentence 1	Word 1	Word 2	Word 3	Metadata 1	Metadata 2	Metadata 3
Sentence 2	Word 1	Word 2	Word 3	Metadata 1	Metadata 2	Metadata 3
Sentence 3	Word 1	Word 2	Word 3	Metadata 1	Metadata 2	Metadata 3



Mapping Example



Impression & Recommendations:

Problem # 1: Physical exam, routine (ICD-V70.0) (ICD10-Z00.00)
Reviewed preventive care protocols, scheduled due services, and updated immunizations.

Orders:
Comp Metabolic (CMP)
Lipid Panel (LPDPA)
CBC No differential (CBCND)
Thyroid Cascade (TSHRC)
Hemoglobin A1C (HA1CG)

Problem # 2: Tics of organic origin (ICD-333.3) (ICD10-G25.69)
Will refer to neurology as pt has neurosurgeon but not neurologist. Discussed with pt having consult for the tics as been present for years. Pt stated didnt improve with shunt. Pt does not feel any medication has worsened it.
Orders:
Neurology Consult & Treat (*)

Problem # 3: Obesity (ICD-278.00) (ICD10-E66.9)
Encouraged cont diet and exercise.
Will start on wellbutrin to see if helpful with weight loss and pt symptoms of what she feels is ADHD.
Discussed returning to weight loss clinic for medication management as she is fearful of sleeve.

Mapping Constraints

- ➔ Identification
- ➔ Position
- ➔ Isolation
- ➔ Page Layout