

Combining NLP and Tabular Features

Mapping Clinical Entities to the relevant Section Headers

Sagar Dawda

About Me



Sagar Dawda

Data Scientist III



- Design and Implement solutions with the help of data
- NLP on large scale on clinical text
- Data Science mentor

RECENT PROJECTS

Entity to
Section
Header
mapping

Breaking
charts into
Clinical
Encounters

Ontology
creation and
Search Engine

About Episource and the DS team

The Company

A Healthcare Risk Management company catering to organizations like insurance providers and hospitals

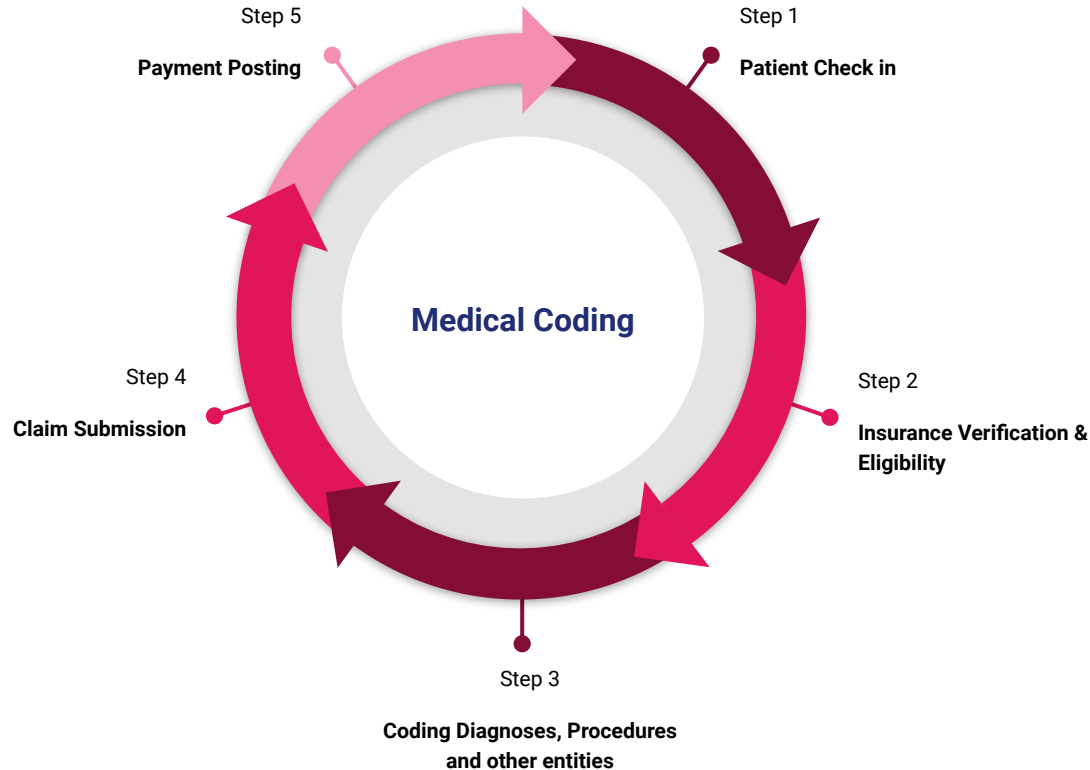
Operate within a strict HIPPA compliant environment to render the services

The Team

Team of 20 strong people including data Scientists, Data Engineers and Subject Matter Experts

Build robust and scalable healthcare solutions with the help of ML & NLP

The cycle of Medical Coding



Refer to the link for more info -

<https://www.m-scribe.com/blog/6-signs-you-have-a-great-medical-billing-process>

Definition, Importance & Objective of the project



Section Headers give a gist of the content that follows and also marks the change in section for a given document

- Determines the relevance of a disease
- Can lead to current state of a disease
- Has a direct impact on the claims / reimbursement

IMPORTANCE

- Identify actual Section Headers
- Map the disease to its relevant header

OBJECTIVES

Example

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**

Review of Systems

See HPI; All other systems are negative on ROS.

Challenges in the project



- No standard chart formats
- Standards available for what to include but not how
- Multiple formats within a single chart







- Disease as Section Headers
- Content of a section spans across multiple pages



- OCR errors
- Text alignment in multiple column layout
- Content of a section spanning across multiple pages




Impact of Section Headers

Accurate and valid disease captured for reimbursement	<p>Correct Section & Mapping for valid disease</p> 	<p>Correct Section & / OR Incorrect Mapping for valid disease</p> 	May lead to under-coding and reimbursement affected
May lead to over-coding and client may face penalty	<p>Incorrect Section & / OR Incorrect Mapping for invalid disease</p> 	<p>Correct Section & Mapping for invalid disease</p> 	Avoids unwanted coding and errors in reimbursement

Initial Approaches

	Sentence Classification Approach	Shortlisted Approach	Metadata Approach
Approach Overview	<ol style="list-style-type: none">1. Tokenization2. Lower casing3. Stemming / Lemmatization4. Stop words removal5. Vectorization6. Model training & testing ML Model	<ol style="list-style-type: none">1. Tokenization2. Lower casing3. Stemming / Lemmatization4. Stop words removal5. Vectorization6. Model training & testing CRF Model	<ol style="list-style-type: none">1. Tabulate feature set2. Convert categories to integer3. Feature selection4. Stop words removal5. Model training & testing
Pros	<ol style="list-style-type: none">1. Using TFIDF rare words gain more importance2. Easy on computation	<ol style="list-style-type: none">1. Gives more importance to sequence in context2. Not sensitive to imbalance	<ol style="list-style-type: none">1. Works with limited set of data2. No need for complex DL models
Cons	<ol style="list-style-type: none">1. Its a BOW approach2. Ignores sequence and semantics	<ol style="list-style-type: none">1. High computational complexity2. Doesnt work well with unknown words	<ol style="list-style-type: none">1. Ignores text content totally2. Feature overlap
Best Use Case	Identify most important keywords	Named Entity Recognition	Any tabular / structured data

Drawback of the existing approach

 1	Section Headers are not sequence problems	<ul style="list-style-type: none">• <u>Chief Complaint</u> - Relevant Section Header• Patient is taking medications for all the conditions listed in <u>chief complaint</u>
 2	Too many Type 1 errors	<ul style="list-style-type: none">• Irrelevant sections identified• Incorrect mappings due to additional headers identified
 3	Text sequence may not match the PDF format	<ul style="list-style-type: none">• This issue majorly occurs with tables• Broken sequence does not work well with CRF approach

Example

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**

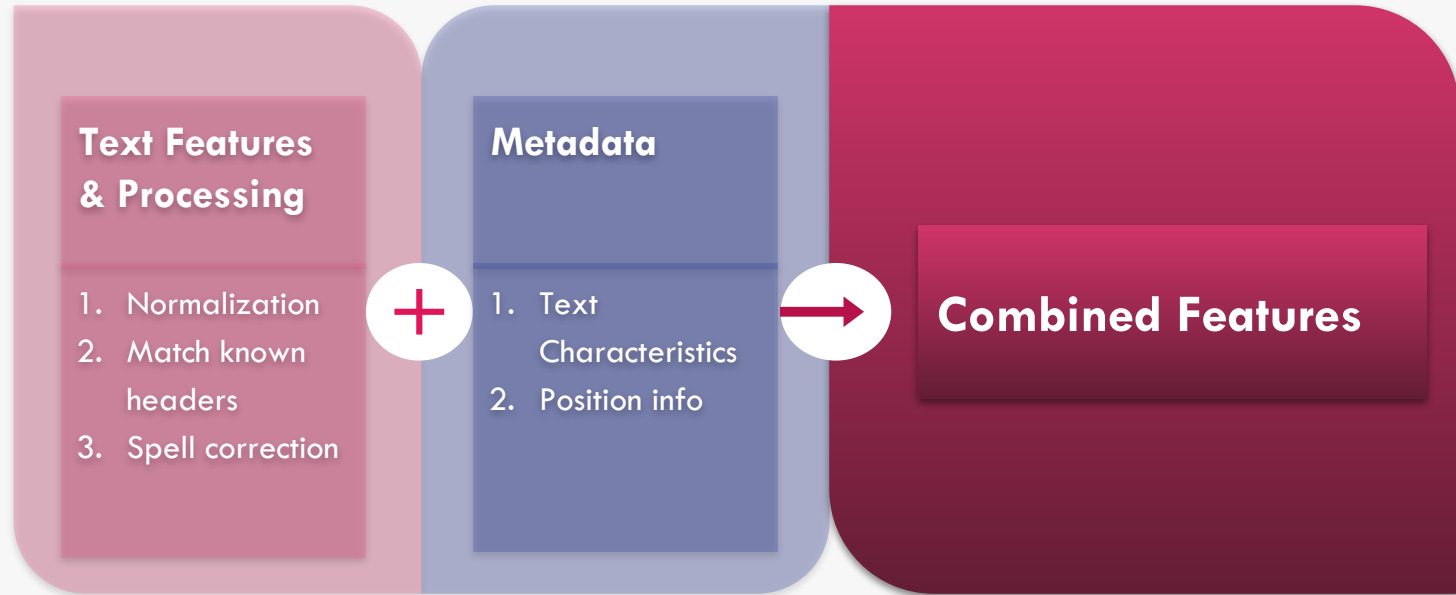
Review of Systems

See **HPI**. All other systems are negative on ROS.

Section
Headers
incorrectly
classified as
Disease

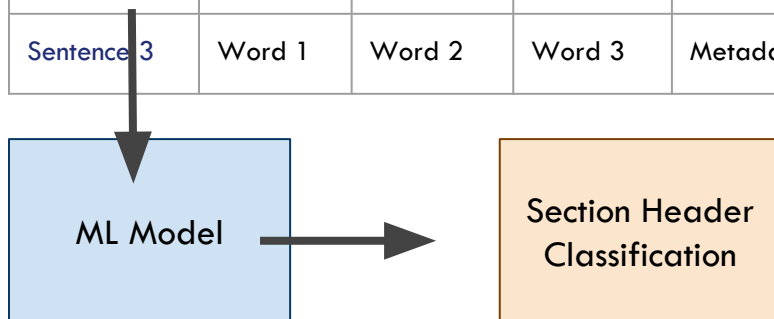
Incorrect
Header

The Solution



Sample of final matrix passed to ML Model

Text ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	... Feature n
Sentence 1	Word 1	Word 2	Word 3	Metadata 1	Metadata 2	Metadata 3	... data n
Sentence 2	Word 1	Word 2	Word 3	Metadata 1	Metadata 2	Metadata 3	... data n
Sentence 3	Word 1	Word 2	Word 3	Metadata 1	Metadata 2	Metadata 3	... data n



Additional Challenges



Memory Efficient
XML Parsing



Processing time per
chart



Complexity vs
Accuracy Tradeoff

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 (*)

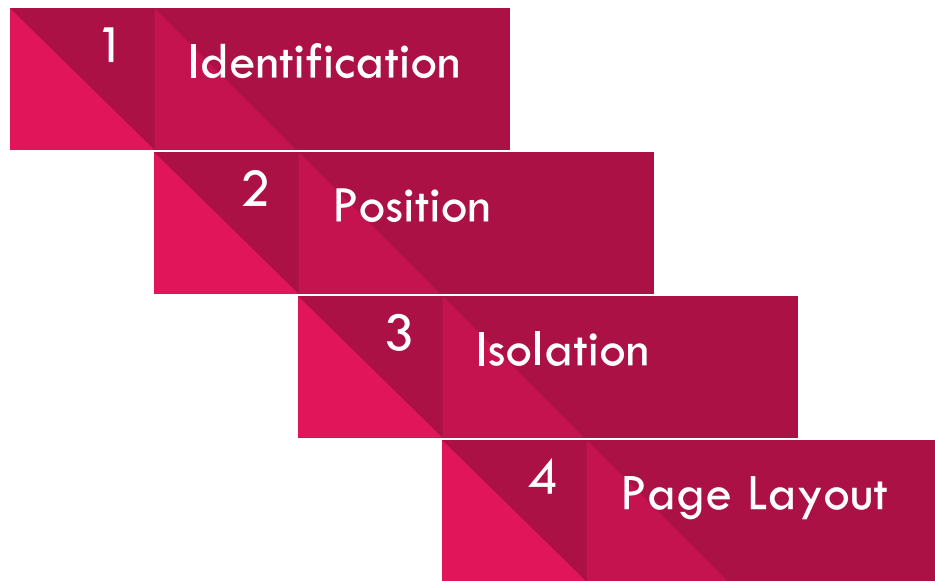
Problem # 3: Obesity (ICD-278.00) (ICD10-E66.9)

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



+19%

	CRF	Hybrid
Mapping Accuracy	67%	86%