Disease Prediction Project Update

April 3, 2023

Progress So Far

- Data Collection
- Data Processing
- Data Paraphrasing for Model Use

Data Collection

3 Main Datasets:

- Kaggle: Diseases and their Symptoms
- Kaggle: Disease Symptom Prediction
- Columbia University: Disease-Symptom Knowledge Database

Diseases and Their Symptoms Dataset

- Contains roughly 800 diseases and their corresponding symptoms
- Original data poorly formatted
  - Inconsistent number of columns and information in each column
  - Required thorough cleaning to be useful

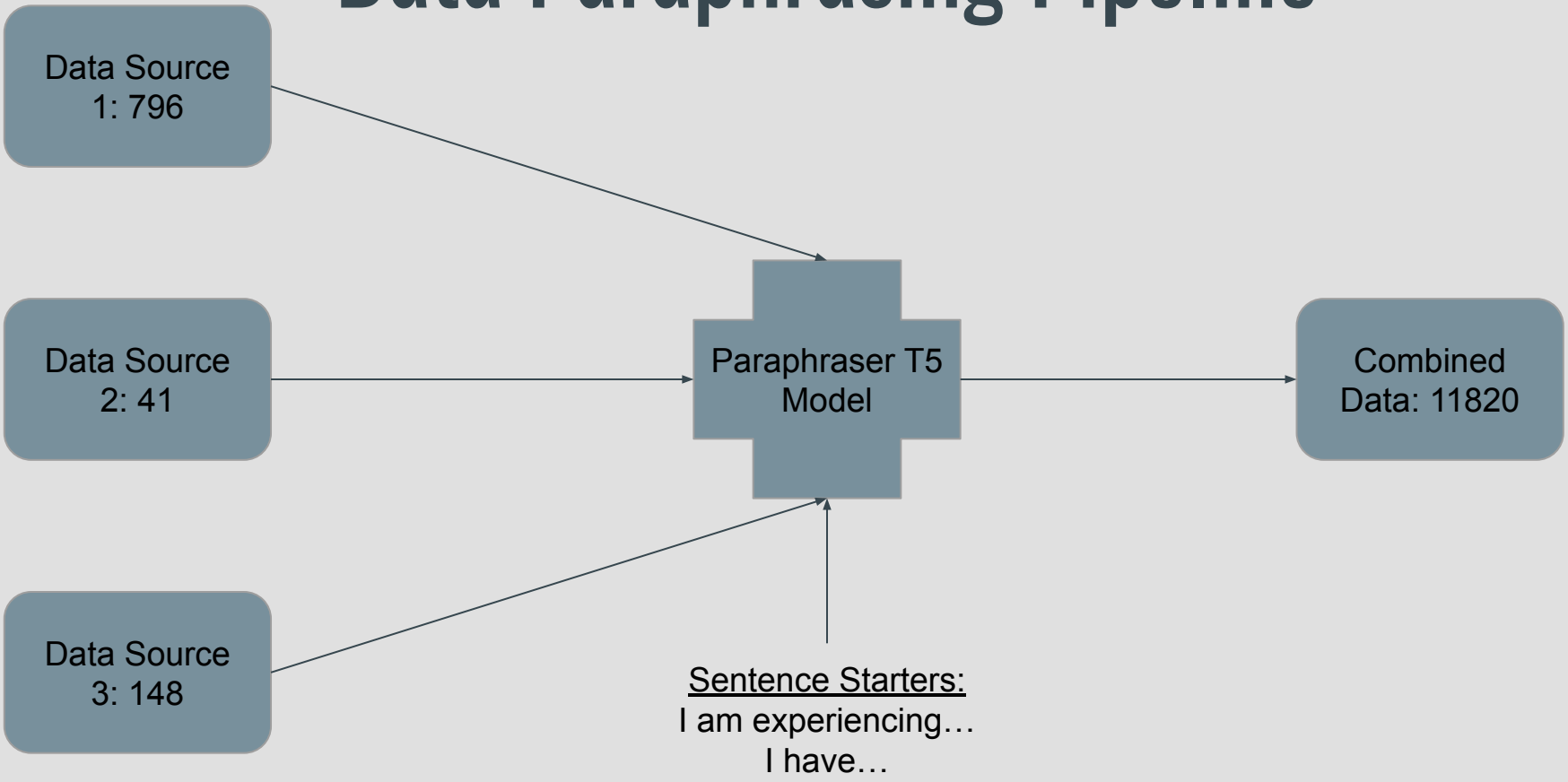
Disease Symptom Prediction Dataset

- Contains roughly 40 diseases and their corresponding symptoms
  - Small amount of data but unique diseases
- Original data also poorly formatted
  - Inconsistent number of columns and information in each column
  - Required thorough cleaning to be useful again

Disease-Symptom Knowledge Database

- Contains roughly 150 diseases and their corresponding symptoms
- Original data in tabular format on website
  - Medical diagnosis code attached to each disease and symptom
  - Required diagnosis code separation and web scraping to be useful

Data Paraphrasing Pipeline



T5 Model: Vamsi/T5_Paraphrase_Paws

Aggregated Data

	Unnamed: 0	Disease	Symptoms
0	0	fungal infection	itching, skin rash, nodal skin eruptions, d...
1	1	allergy	continuous sneezing, shivering, chills, wa...
2	2	gerd	stomach pain, acidity, ulcers on tongue, v...
3	3	chronic cholestasis	itching, vomiting, yellowish skin, nausea, ...
4	4	drug reaction	itching, skin rash, stomach pain, burning m...
5	5	peptic ulcer diseae	vomiting, loss of appetite, abdominal pain,...
6	6	aids	muscle wasting, patches in throat, high fev...
7	7	diabetes	fatigue, weight loss, restlessness, lethar...
8	8	gastroenteritis	vomiting, sunken eyes, dehydration, diarrhoea
9	9	bronchial asthma	fatigue, cough, high fever, breathlessness...
10	10	hypertension	headache, chest pain, dizziness, loss of b...
11	11	migraine	acidity, indigestion, headache, blurred an...
12	12	cervical spondylosis	back pain, weakness in limbs, neck pain, d...
13	13	paralysis	vomiting, headache, weakness of one body si...
14	14	jaundice	itching, vomiting, fatigue, weight loss, h...

Paraphrased Data

	Unnamed: 0	Disease	Symptoms
0	0	fungal infection	I have itching, skin rash, nodal skin erupti...
1	1	allergy	I have continuous sneezing, shivering, chil...
2	2	gerd	I have stomach pain, acidity, ulcers on ton...
3	3	chronic cholestasis	I have itching, vomiting, yellowish skin, n...
4	4	drug reaction	I have itching, skin rash, stomach pain, bu...
5	5	peptic ulcer diseae	I have vomiting, loss of appetite, abdomina...
6	6	aids	I have muscle wasting, patches in throat, h...
7	7	diabetes	I have fatigue, weight loss, restlessness, ...
8	8	gastroenteritis	I have vomiting, sunken eyes, dehydration, ...
9	9	bronchial asthma	I have fatigue, cough, high fever, breathl...
10	10	hypertension	I have headache, chest pain, dizziness, lo...
11	11	migraine	I have acidity, indigestion, headache, blu...
12	12	cervical spondylosis	I have back pain, weakness in limbs, neck p...
13	13	paralysis	I have vomiting, headache, weakness of one ...
14	14	jaundice	I have itching, vomiting, fatigue, weight l...

Next Steps

- Split data into training/testing batches
- Fine-tune the chosen model (CORe Clinical Diagnosis Prediction Model) with curated data
- Get more data as needed
- Develop secondary model for patient Q&A or treatment recommendation???