t-SNE Visualization for MIMIC IV Medical Notes

Sonu Sud (ss232867)

1. Extracting Entities using SpaCy and SciSpacy

We will use both spaCy and SciSpacy to extract named entities from the medical notes. SciSpacy is specifically designed for biomedical and clinical text.

Steps:

1. Load spacy and scispacy small core models

```
# Load spaCy and SciSpacy models
spacy_nlp = spacy.load('en_core_web_sm')
scispacy_nlp = spacy.load('en_core_sci_sm')
```

2. Extract named entities using spacy_nlp model in spaCy
3. Extract named entities using scispacy_nlp model in sciSpacy

Sample Entities

```
/opt/miniconda3/envs/mimic-vis/lib/python3.11/site-packages/spacy/language.py:2195: FutureWarning: Possible set union at position 6328
  deserializers["tokenizer"] = lambda p: self.tokenizer.from_disk(  # type: ignore[union-attr]
spaCy Entities:
[('Chief Complaint', 'PERSON'), ('O2', 'NORP'), ('95%', 'PERCENT'), ('6L', 'CARDINAL'), ('O2', 'CARDINAL'), ('102.3', 'CARDINAL'),
('ED', 'GPE'), ('CXR', 'ORG'), ('RLL', 'ORG'), ('UA', 'ORG'), ('SBP', 'ORG'), ('80', 'CARDINAL'), ('up to 100', 'CARDINAL'),
('1L', 'CARDINAL'), ('NS', 'GPE'), ('110', 'CARDINAL'), ('MICU', 'ORG'), ('DNR', 'ORG'), ('T99', 'PRODUCT'), ('O2 94%',
'PERCENT'), ('10L.', 'DATE'), ('CHF', 'ORG'), ('Family History', 'PERSON'), ('ADMISSION', 'ORG'), ('General-', 'GPE'), ('Elderly',
'GPE'), ('CV-', 'PERSON'), ('Tachycardic', 'PERSON'), ('Voluntary', 'PERSON'), ('TTP', 'ORG'), ('Neuro', 'PERSON'), ('MAE',
'ORG'), ('DISCHARGE', 'ORG'), ('Pertinent Results', 'PERSON'), ('BLOOD WBC-7.0', 'ORG'), ('PTT-28.5', 'PERSON'), ('# \nNa-149',
'ORG'), ('29', 'CARDINAL'), ('AnGap-17', 'NORP'), ('02:21AM BLOOD', 'PERSON'), ('06:45PM', 'PRODUCT'), ('12:55AM', 'CARDINAL'),
('06:50PM', 'PERSON'), ('pO2-35', 'ORG'), ('calTCO2-34', 'DATE'), ('06:50PM', 'PERSON'), ('CHF', 'ORG'), ('MRSA PNA', 'ORG'),
('UTI', 'ORG'), ('CMO', 'ORG'), ('Autopsy', 'ORG'), ('ISSUES', 'ORG'), ('Pneumonia', 'MONEY'), ('CXR', 'ORG'), ('RLL', 'ORG'),
('ED', 'GPE'), ('HCAP', 'ORG'), ('vanc/levofloxacin', 'ORG'), ('CXR', 'ORG'), ('O2', 'CARDINAL'), ('CMO', 'ORG'), ('Autopsy',
'ORG'), ('#', 'CARDINAL'), ('Hypercarbic', 'ORG'), ('NC', 'GPE'), ('SNF', 'ORG'), ('O2', 'CARDINAL'), ('VBG', 'ORG'), ('7',
'CARDINAL'), ('UTI', 'ORG'), ('UA', 'ORG'), ('86RBC', 'ORG'), ('182', 'CARDINAL'), ('WBC', 'ORG'), ('Enterococcus', 'ORG'), ('#
MSSA Bacteremia', 'MONEY'), ('#Altered Mental Status', 'MONEY'), ('Haldol', 'ORG'), ('DM', 'MONEY'), ('10', 'DATE'), ('SNF',
'ORG'), ('ISS', 'ORG'), ('#Humerus Fracture', 'ORG'), ('CKD', 'ORG'), ('Baseline 1.7', 'PERSON'), ('3.7', 'CARDINAL'), ('AMS',
'ORG'), ('PRN', 'ORG'), ('50-55%', 'PERCENT'), ('40', 'CARDINAL'), ('PO BID', 'FAC'), ('CAD', 'ORG'), ('Glaucoma', 'GPE'), ('PO',
'ORG'), ('Autopsy', 'ORG'), ('Medications', 'PERSON'), ('1', 'CARDINAL'), ('10', 'CARDINAL'), ('PRN', 'ORG'), ('2', 'CARDINAL'),
('100', 'CARDINAL'), ('PO BID', 'FAC'), ('3', 'CARDINAL'), ('650', 'CARDINAL'), ('PO', 'GPE'), ('4', 'CARDINAL'), ('Senna',
'PERSON'), ('8.6', 'CARDINAL'), ('PO BID', 'FAC'), ('PRN', 'ORG'), ('5', 'CARDINAL'), ('PO DAILY', 'FAC'), ('6', 'CARDINAL'), ('PO
DAILY', 'FAC'), ('7', 'CARDINAL'), ('PO DAILY', 'FAC'), ('8', 'CARDINAL'), ('PO DAILY', 'ORG'), ('9', 'CARDINAL'), ('PrednisoLONE
Acetate 1% Ophth', 'ORG'), ('1', 'CARDINAL'), ('10', 'CARDINAL'), ('0.005%', 'PERCENT'), ('1', 'CARDINAL'), ('11', 'CARDINAL'),
('Dorzolamide 2%/Timolol', 'ORG'), ('0.5%', 'PERCENT'), ('Ophth', 'ORG'), ('1', 'CARDINAL'), ('12', 'CARDINAL'), ('40',
'CARDINAL'), ('PO BID', 'FAC'), ('13', 'CARDINAL'), ('PO BID', 'ORG'), ('14', 'CARDINAL'), ('12.5', 'CARDINAL'), ('PO BID',
'FAC'), ('15', 'CARDINAL'), ('0.5 %', 'PERCENT'), ('daily', 'DATE'), ('16', 'CARDINAL'), ('Vancomycin 25mg/mL Ophth Soln 1',
'ORG'), ('LEFT EYE DAILY', 'PERSON'), ('17', 'CARDINAL'), ('12', 'CARDINAL'), ('18', 'CARDINAL'), ('AILY', 'PERSON'), ('PRN',
'ORG'), ('19', 'CARDINAL'), ('1', 'CARDINAL'), ('Discharge Medications', 'PERSON'), ('Expired\n \nDischarge Disposition', 'ORG'),
('Expired\n \nDischarge Diagnosis', 'ORG'), ('MSSA Bacteremia\nEnterococcus UTI\n\n \nDischarge Condition', 'WORK_OF_ART'),
('Expired\n \nFollowup Instructions', 'ORG')]

SciSpacy Entities:
[('Admission', 'ENTITY'), ('Discharge', 'ENTITY'), ('Sex', 'ENTITY'), ('Atenolol', 'ENTITY'), ('Attending', 'ENTITY'), ('Chief
Complaint', 'ENTITY'), ('Respiratory distress', 'ENTITY'), ('Surgical', 'ENTITY'), ('Invasive Procedure', 'ENTITY'), ('history',
'ENTITY'), ('COPD', 'ENTITY'), ('6L home O2', 'ENTITY'), ('prior', 'ENTITY'), ('respiratory \nfailure', 'ENTITY'), ('IDDM',
'ENTITY'), ('dementia', 'ENTITY'), ('ED', 'ENTITY'), ('humeral fracture', 'ENTITY'), ('nursing home', 'ENTITY'), ('respiratory
distress', 'ENTITY'), ('ED', 'ENTITY'), ('AOx2', 'ENTITY'), ('baseline', 'ENTITY'), ('exam', 'ENTITY'), ('diffuse rhonci',
'ENTITY'), ('satting', 'ENTITY'), ('6L', 'ENTITY'), ('baseline home O2', 'ENTITY'), ('Fever', 'ENTITY'), ('ED', 'ENTITY'), ('CXR',
'ENTITY'), ('RLL infiltrate', 'ENTITY'), ('effusion', 'ENTITY'), ('UA', 'ENTITY'), ('positive', 'ENTITY'), ('vancomycin',
'ENTITY'), ('ceftriaxone', 'ENTITY'), ('azithromycin', 'ENTITY'), ('SBP', 'ENTITY'), ('dipped', 'ENTITY'), ('briefly', 'ENTITY'),
('went', 'ENTITY'), ('NS', 'ENTITY'), ('HR', 'ENTITY'), ('head CT', 'ENTITY'), ('ED', 'ENTITY'), ('negative', 'ENTITY'), ('bleed',
'ENTITY'), ('admitted', 'ENTITY'), ('MICU', 'ENTITY'), ('respiratory \ndistress', 'ENTITY'), ('intubated', 'ENTITY'),
('discharge', 'ENTITY'), ('admission', 'ENTITY'), ('DNR/DNI', 'ENTITY'), ('nursing', 'ENTITY'), ('ED', 'ENTITY'), ('nursing',
'ENTITY'), ('code', 'ENTITY'), ('arrival', 'ENTITY'), ('floor', 'ENTITY'), ('moaning', 'ENTITY'), ('movement', 'ENTITY'),
('Vitals', 'ENTITY'), ('transfer \n', 'ENTITY'), ('HR92 BP106/48 RR26 O2', 'ENTITY'), ('Past Medical History', 'ENTITY'),
('corneal inplant', 'ENTITY'), ('L eye', 'ENTITY'), ('glaucoma', 'ENTITY'), ('CAD', 'ENTITY'), ('prostate cancer', 'ENTITY'),
('anemia', 'ENTITY'), ('depression', 'ENTITY'), ('arthritis', 'ENTITY'), ('HTN', 'ENTITY'), ('respiratory failure', 'ENTITY'),
('past', 'ENTITY'), ('wound)\n', 'ENTITY'), ('dementia', 'ENTITY'), ('CHF', 'ENTITY'), ('Social History', 'ENTITY'), ('Family
History', 'ENTITY'), ('DM', 'ENTITY'), ('father', 'ENTITY'), ('mother', 'ENTITY'), ('No cancers', 'ENTITY'), ('Physical Exam',
'ENTITY'), ('Elderly', 'ENTITY'), ('ill appearing', 'ENTITY'), ('moaning', 'ENTITY'), ('HEENT- L', 'ENTITY'), ('dialated',
'ENTITY'), ('corneal implant', 'ENTITY'), ('sclera \n', 'ENTITY'), ('MMM', 'ENTITY'), ('R pupil reactive', 'ENTITY')…….
```

2. Train Word2Vec using spaCy and sciSpacy Entities

We will train a Word2Vec model on the medical notes and use t-SNE to visualize the embeddings of the extracted entities.

Steps:

1. Train Word2Vec model for each of spacy and scispacy entities corpus to generate embeddings

```
w2v_spacy = Word2Vec(spacy_corpus, window=5, min_count=1, workers=4)
w2v_scispacy = Word2Vec(scispacy_corpus, window=5, min_count=1, workers=4)
```

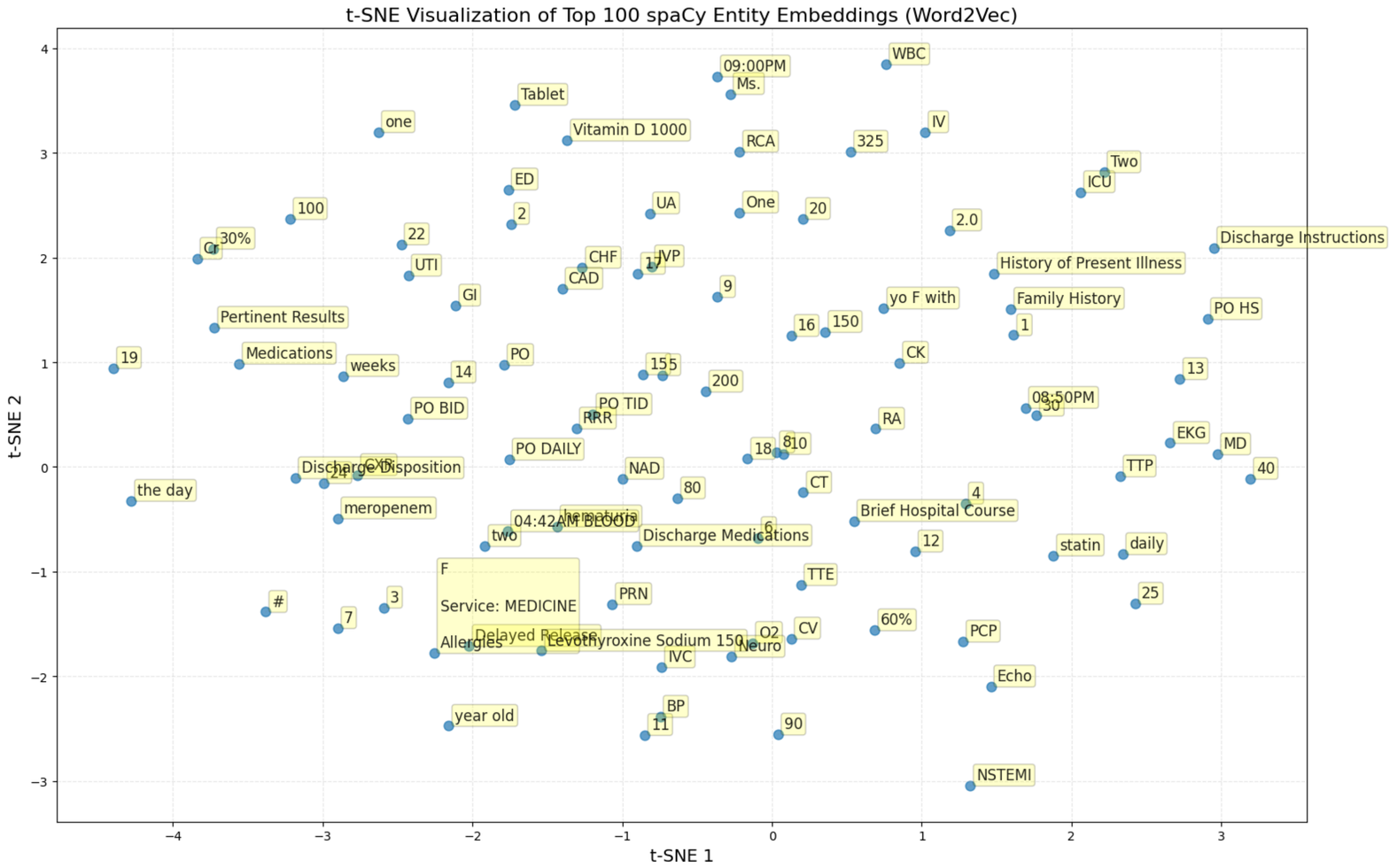
2. For Readability, filter top 100 most common or frequent entities to build the t-SNE plot
3. I used used dynamic perplexity calculation for t-SNE to handle cases where sample size is lower than 30
4. Used PCA for projecting the multi dimensional word2Vec tensor to 2 components

3. Word2Vec Embedding and t-SNE Visualization on SpaCy Entities

Data Used: Discharge Notes from MIMIC IV Notes Dataset for 100 Patients

Interpretation:
This plot shows how the most frequent entities extracted by the general-purpose spaCy model are distributed in semantic space, based on their Word2Vec embeddings.

- Clusters: Entities that appear close together are used in similar contexts in the clinical notes, suggesting semantic similarity.
- Spread: If the plot shows tight clusters, it means spaCy is grouping related medical concepts well. If it's more scattered, spaCy may be capturing a broader, less domain-specific vocabulary.
- Labels: The most frequent entities are likely to be general medical terms, patient demographics, or common clinical concepts.
- Due to generic nature of entities, The SpaCy data does not provide any meaningful information

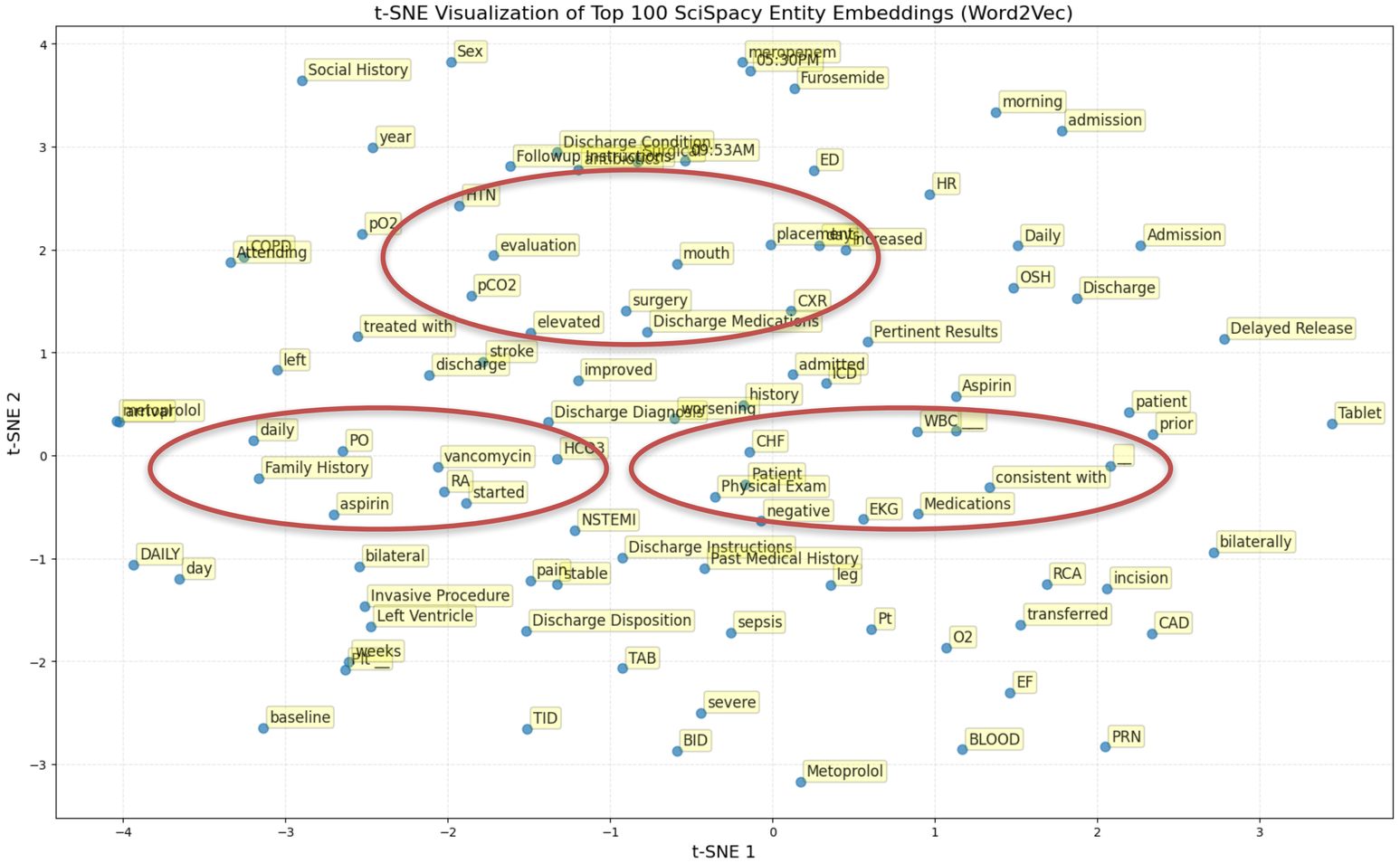


4. Word2Vec Embedding and t-SNE Visualization on SciSpaCy Entities

Data Used: Discharge Notes from MIMIC IV Notes Dataset for 100 Patients

Interpretation:
This plot visualizes the semantic relationships between entities extracted by SciSpacy, a model specialized for biomedical and clinical text.

- Clusters: Expect tighter, more meaningful clusters of clinical concepts (e.g., diseases, medications, procedures) compared to spaCy.
- Domain Focus: SciSpacy's domain knowledge should result in more clinically relevant groupings, helping to identify related medical entities and terminology.
- Labels: The entities are more likely to be technical or specific to healthcare, reflecting the model's biomedical focus.
- Related terms such as medications, discharge information, procedures appear closer together.

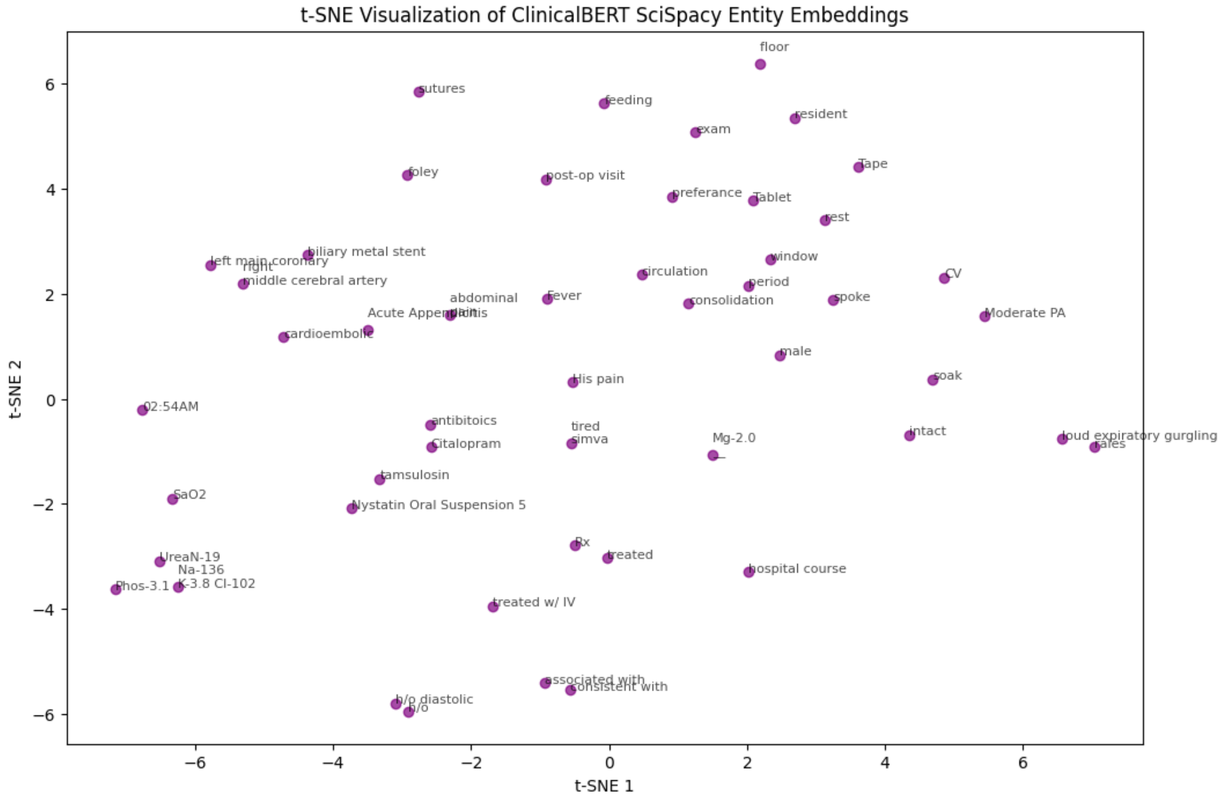


5. t-SNE Visualization on SciSpaCy Entities using ClinicalBERT Embeddings

we will use ClinicalBERT, a transformer-based model pre-trained on clinical notes, to generate embeddings for medical entities and visualize them with t-SNE.

Interpretation:
This plot shows how SciSpacy entities are embedded using ClinicalBERT, a transformer model pre-trained on clinical notes.

- Deep Context: ClinicalBERT captures richer, context-aware representations, so clusters may reflect deeper semantic relationships (e.g., grouping symptoms with related diagnoses).
- Clinical Relevance: Entities that cluster together are likely to co-occur in similar clinical scenarios or patient cases.
- Labels: The plot highlights which clinical concepts are semantically close according to ClinicalBERT, potentially revealing relationships not captured by Word2Vec.
- ClinicalBERT clearly outperforms Word2Vec as related terms such as chemicals, procedures and medications are represented in close clusters



Steps:

1. Load ‘emilyalsentzer/Bio_ClinicalBERT' from hugging face
2. Generate ClinicalBERT embeddings for scispacy entities
3. I used dynamic perplexity calculation for t-SNE to handle cases where sample size is lower than 30