MIMIC-III - NLP

Cole Stokes

AI in Healthcare

The University of Texas at Austin

Google Colab - GitHub

DESCRIPTION

This presentation compares the ability of three libraries/models to recognize and extract named entities from medical notes and visualize them through tSNE plots. The medical notes from MIMIC-III are filtered down to include only notes from patients with Neuromyelitis Optica Spectrum Disorder or NMOSD (ICD Code: 3410).

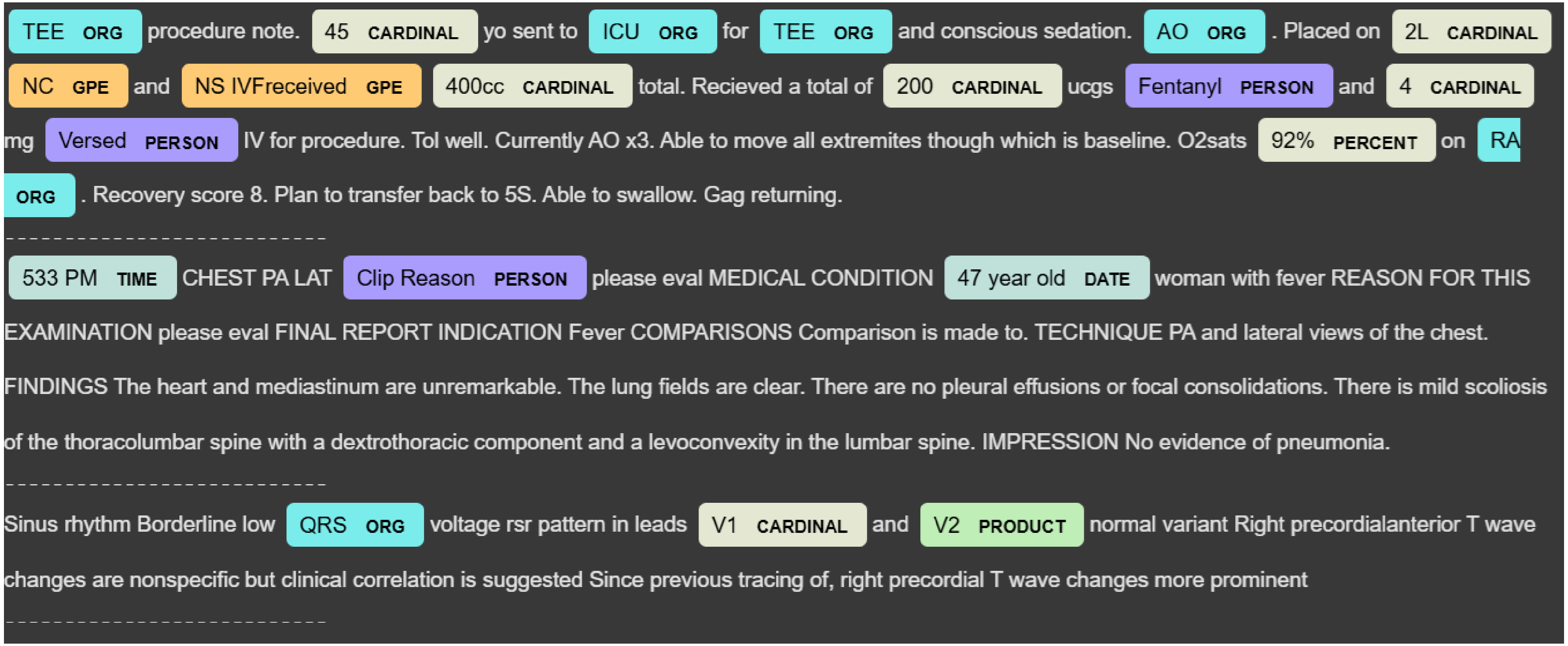
- Entities were extracted using Spacy, Scispacy, and blueBERT.
- For the Spacy and Scispacy models, Spacy's Displacy function visualized recognized entities, and gensim was used to create word2vecs.
- For each model tSNE plots were created using matplotlib and sklearn for both GloVe 50D (a pre-trained model) and the medical notes for the NMOSD patients.

Note: This dataset is only limited to patients admitted to one hospital, the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

PREPROCESSING

- To get the medical notes for the NMOSD patients, a query was sent to the noteevents table to get the text and was joined with the diagnoses_icd table for patients with the 3410 icd9 code.
- Rows are filtered to only contain unique text (duplicate medical notes are removed) to save computation time.
- The text is cleaned by removing excess whitespace and uncommon characters. The stamps on the radiology notes are also filtered out using latex.

SPACY ENTITY VISUALIZATION



TEE ORG procedure note. 45 CARDINAL yo sent to ICU ORG for TEE ORG and conscious sedation. AO ORG . Placed on 2L CARDINAL NC GPE and NS IVFreceived GPE 400cc CARDINAL total. Recieved a total of 200 CARDINAL ucgs Fentanyl PERSON and 4 CARDINAL mg Versed PERSON IV for procedure. Tol well. Currently AO x3. Able to move all extremites though which is baseline. O2sats 92% PERCENT on RA ORG . Recovery score 8. Plan to transfer back to 5S. Able to swallow. Gag returning.

---------------------------

533 PM TIME CHEST PA LAT Clip Reason PERSON please eval MEDICAL CONDITION 47 year old DATE woman with fever REASON FOR THIS EXAMINATION please eval FINAL REPORT INDICATION Fever COMPARISONS Comparison is made to. TECHNIQUE PA and lateral views of the chest. FINDINGS The heart and mediastinum are unremarkable. The lung fields are clear. There are no pleural effusions or focal consolidations. There is mild scoliosis of the thoracolumbar spine with a dextrothoracic component and a levoconvexity in the lumbar spine. IMPRESSION No evidence of pneumonia.

---------------------------

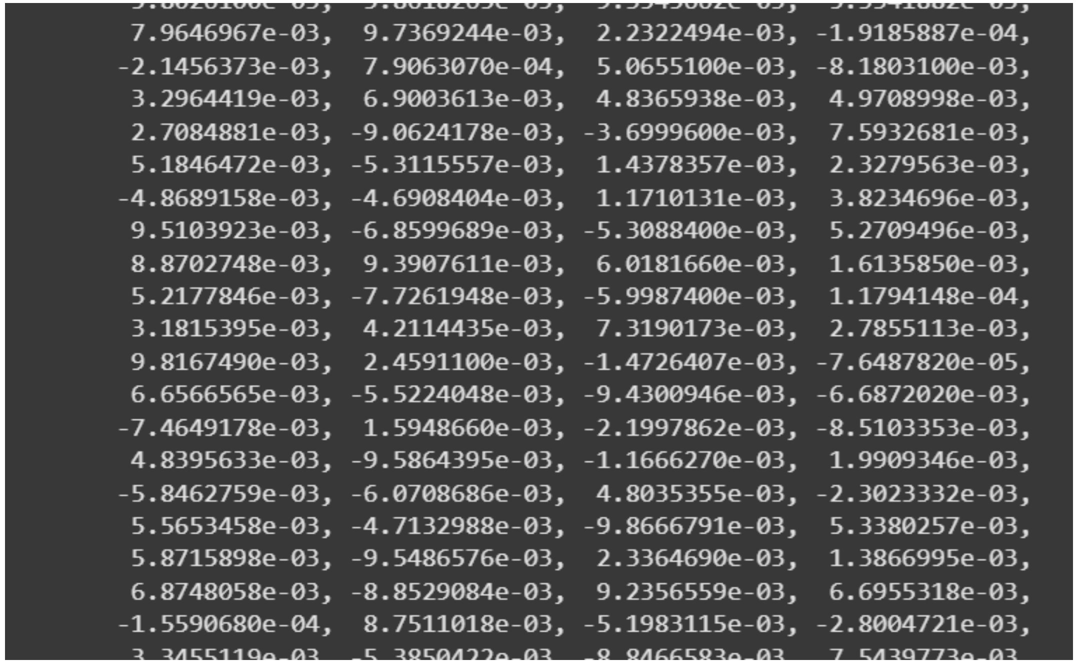
Sinus rhythm Borderline low QRS ORG voltage rsr pattern in leads V1 CARDINAL and V2 PRODUCT normal variant Right precordialanterior T wave changes are nonspecific but clinical correlation is suggested Since previous tracing of, right precordial T wave changes more prominent

---------------------------

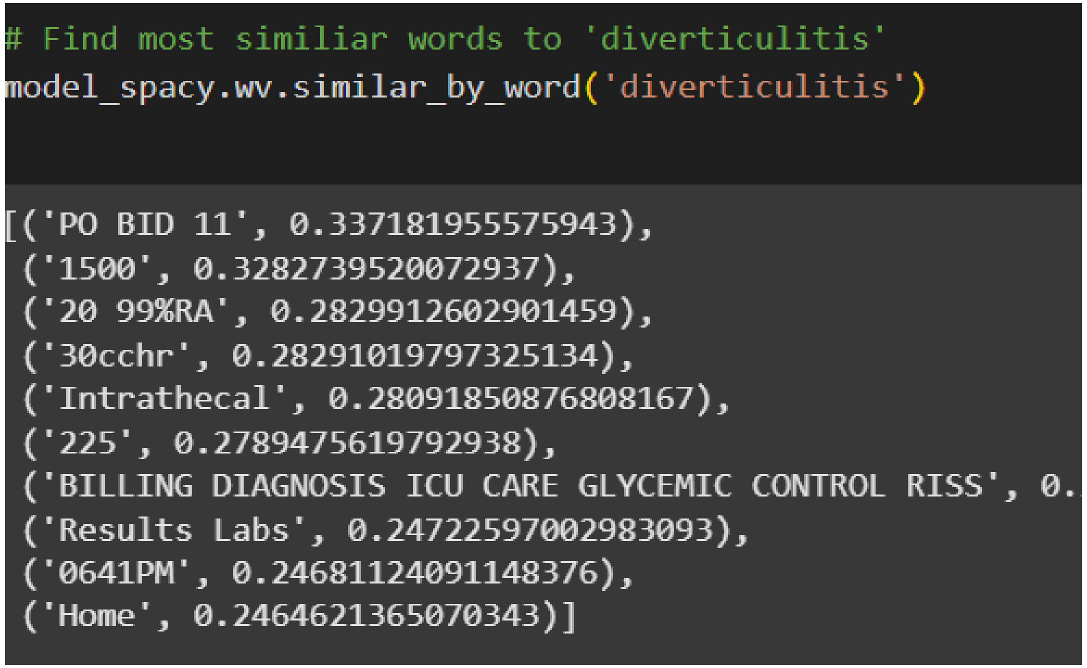
We can see that this doesn't recognize many medical entities or misclassifies them.

SPACY WORD2VEC

The word2vec does not strongly map similar entities to diverticulitis.

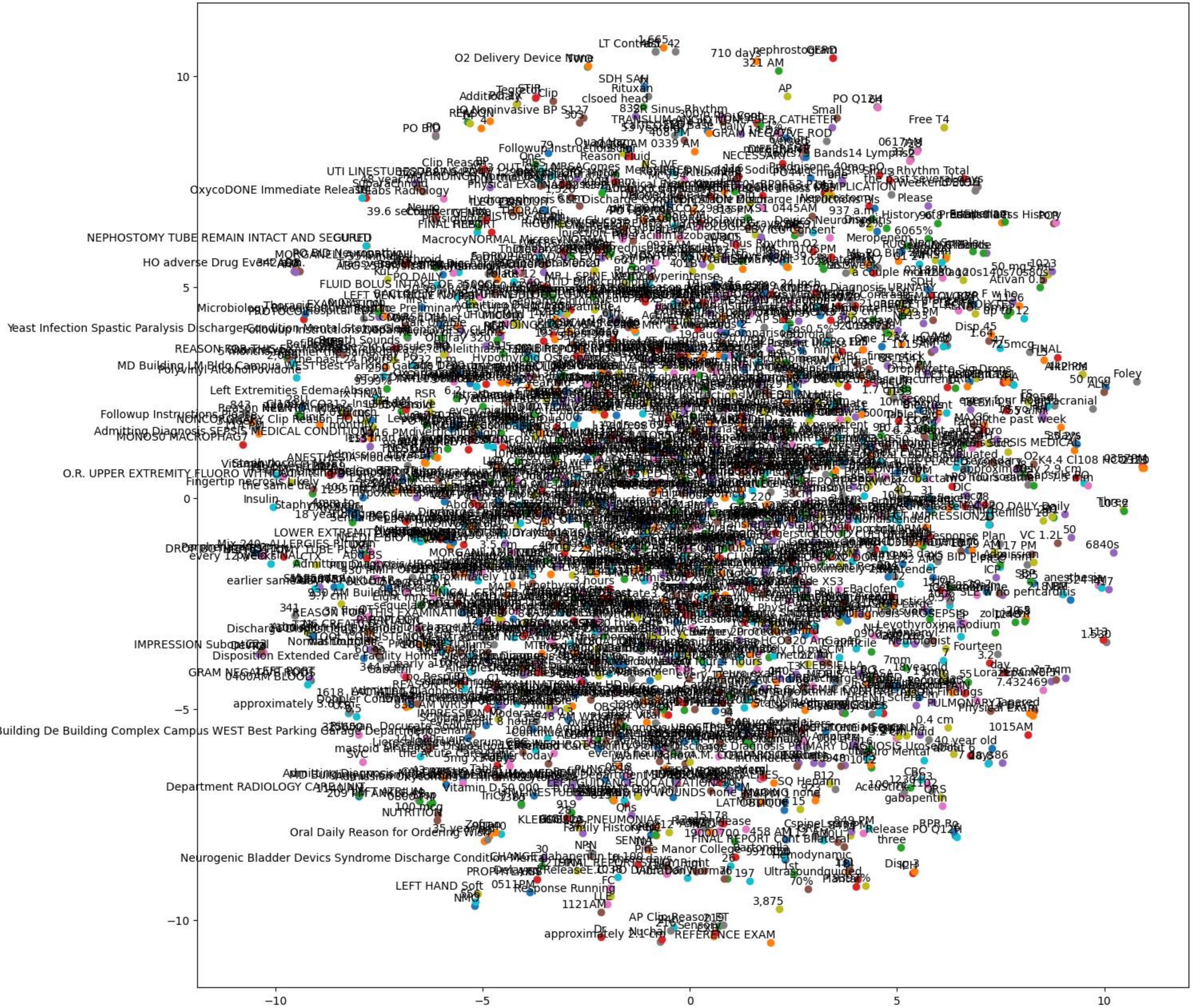


```
 7.9646967e-03,  9.7369244e-03,  2.2322494e-03, -1.9185887e-04,
-2.1456373e-03,  7.9063070e-04,  5.0655100e-03, -8.1803100e-03,
 3.2964419e-03,  6.9003613e-03,  4.8365938e-03,  4.9708998e-03,
 2.7084881e-03, -9.0624178e-03, -3.6999600e-03,  7.5932681e-03,
 5.1846472e-03, -5.3115557e-03,  1.4378357e-03,  2.3279563e-03,
-4.8689158e-03, -4.6908404e-03,  1.1710131e-03,  3.8234696e-03,
 9.5103923e-03, -6.8599689e-03, -5.3088400e-03,  5.2709496e-03,
 8.8702748e-03,  9.3907611e-03,  6.0181660e-03,  1.6135850e-03,
 5.2177846e-03, -7.7261948e-03, -5.9987400e-03,  1.1794148e-04,
 3.1815395e-03,  4.2114435e-03,  7.3190173e-03,  2.7855113e-03,
 9.8167490e-03,  2.4591100e-03, -1.4726407e-03, -7.6487820e-05,
 6.6566565e-03, -5.5224048e-03, -9.4300946e-03, -6.6872020e-03,
-7.4649178e-03,  1.5948660e-03, -2.1997862e-03, -8.5103353e-03,
 4.8395633e-03, -9.5864395e-03, -1.1666270e-03,  1.9909346e-03,
-5.8462759e-03, -6.0708686e-03,  4.8035355e-03, -2.3023332e-03,
 5.5653458e-03, -4.7132988e-03, -9.8666791e-03,  5.3380257e-03,
 5.8715898e-03, -9.5486576e-03,  2.3364690e-03,  1.3866995e-03,
 6.8748058e-03, -8.8529084e-03,  9.2356559e-03,  6.6955318e-03,
-1.5590680e-04,  8.7511018e-03, -5.1983115e-03, -2.8004721e-03,
 3.3455119e-03, -5.3850422e-03, -8.8466583e-03,  7.5439773e-03
```



```
# Find most similiar words to 'diverticulitis'
model_spacy.wv.similar_by_word('diverticulitis')
```

```
[('PO BID 11', 0.337181955575943),
 ('1500', 0.3282739520072937),
 ('20 99%RA', 0.2829912602901459),
 ('30cchr', 0.28291019797325134),
 ('Intrathecal', 0.28091850876808167),
 ('225', 0.2789475619792938),
 ('BILLING DIAGNOSIS ICU CARE GLYCEMIC CONTROL RISS', 0.
 ('Results Labs', 0.24722597002983093),
 ('0641PM', 0.24681124091148376),
 ('Home', 0.2464621365070343)]
```

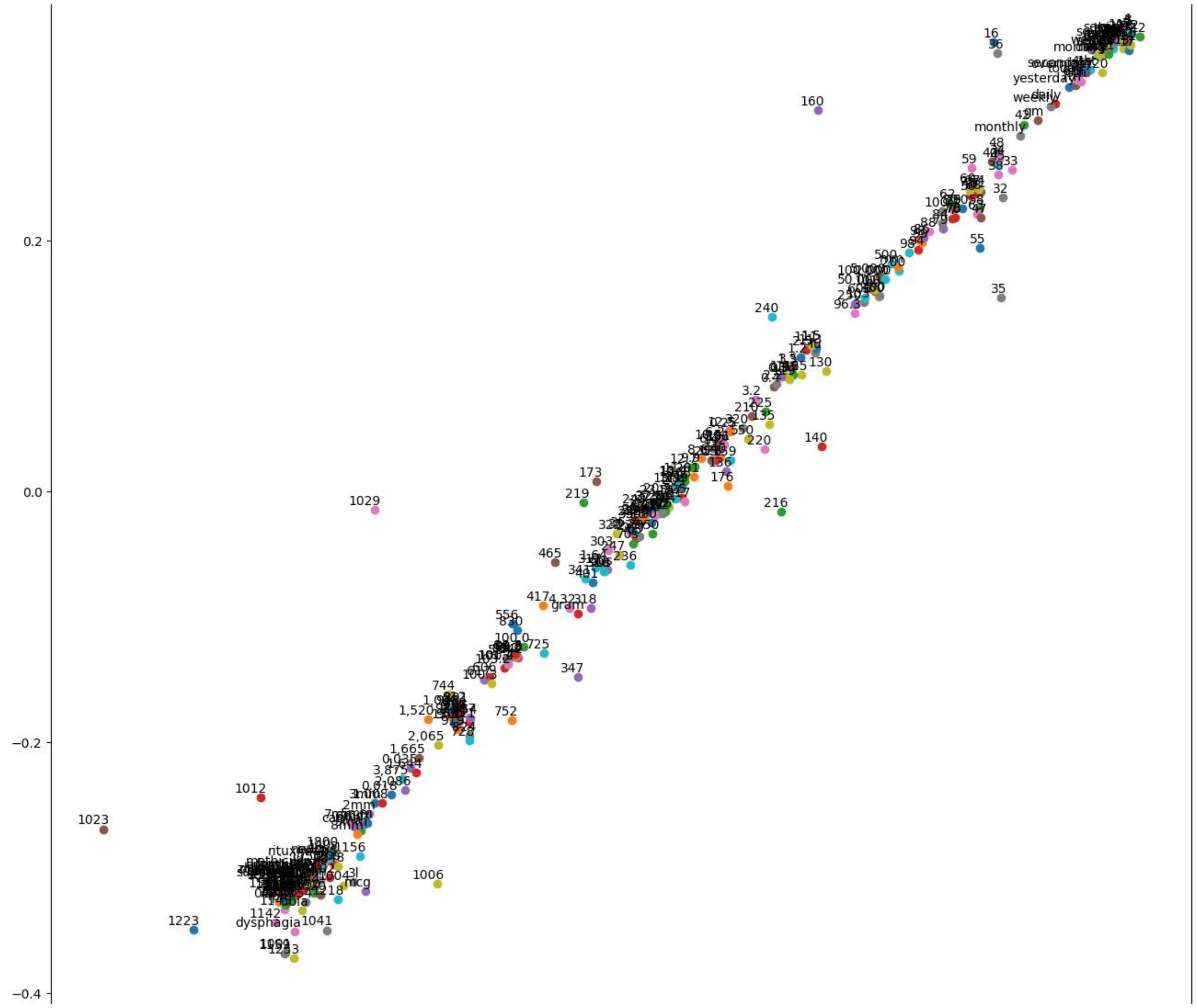


SPACY TSNE PLOT

This Spacy tSNE plot does not seem to clearly map the similarity of the entities.

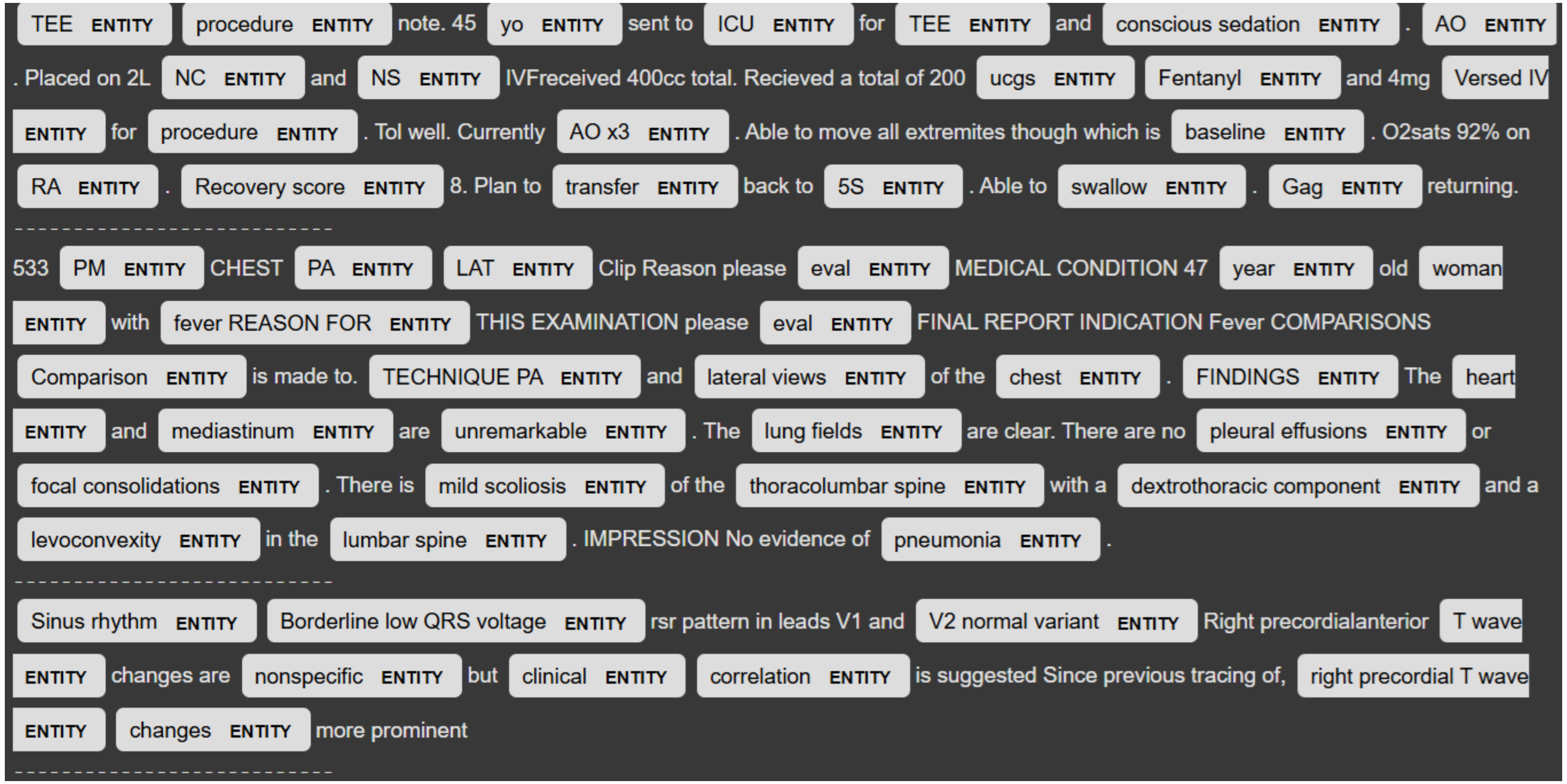
SPACY TSNE PLOT WITH PRE-TRAINED MODEL

This plot shows a bit of an improvement in showing entity similarity.



SCISPACY ENTITY VISUALIZATION

Scipacy seems to pick up a lot more entities than spacy and they seem more relevant to the medical field.



TEE ENTITY procedure ENTITY note. 45 yo ENTITY sent to ICU ENTITY for TEE ENTITY and conscious sedation ENTITY . AO ENTITY . Placed on 2L NC ENTITY and NS ENTITY IVFreceived 400cc total. Recieved a total of 200 ucgs ENTITY Fentanyl ENTITY and 4mg Versed IV ENTITY for procedure ENTITY . Tol well. Currently AO x3 ENTITY . Able to move all extremites though which is baseline ENTITY . O2sats 92% on RA ENTITY . Recovery score ENTITY 8. Plan to transfer ENTITY back to 5S ENTITY . Able to swallow ENTITY . Gag ENTITY returning.

----

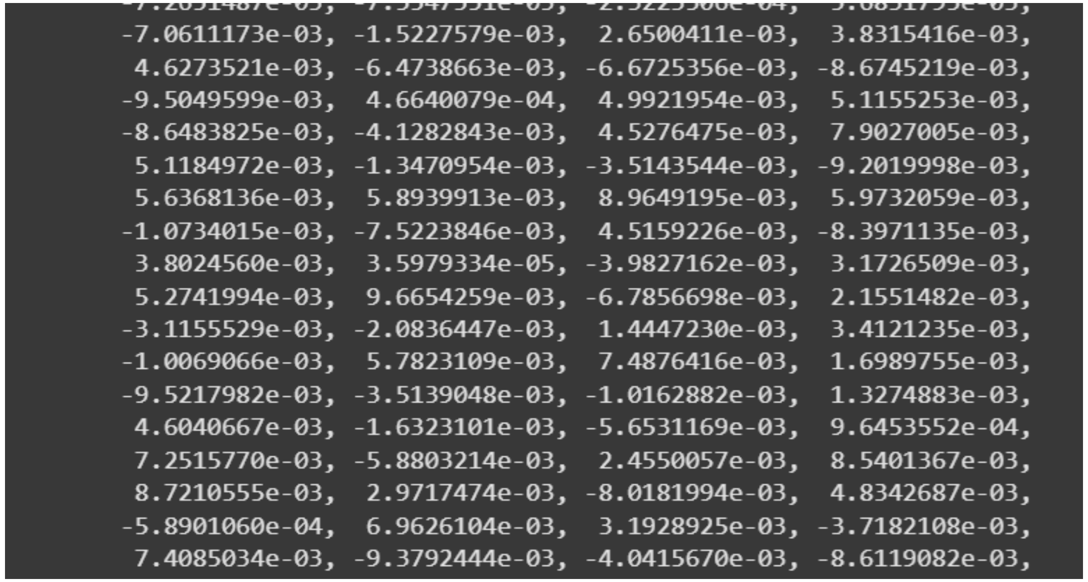
533 PM ENTITY CHEST PA ENTITY LAT ENTITY Clip Reason please eval ENTITY MEDICAL CONDITION 47 year ENTITY old woman ENTITY with fever REASON FOR ENTITY THIS EXAMINATION please eval ENTITY FINAL REPORT INDICATION Fever COMPARISONS Comparison ENTITY is made to. TECHNIQUE PA ENTITY and lateral views ENTITY of the chest ENTITY . FINDINGS ENTITY The heart ENTITY and mediastinum ENTITY are unremarkable ENTITY . The lung fields ENTITY are clear. There are no pleural effusions ENTITY or focal consolidations ENTITY . There is mild scoliosis ENTITY of the thoracolumbar spine ENTITY with a dextrothoracic component ENTITY and a levoconvexity ENTITY in the lumbar spine ENTITY . IMPRESSION No evidence of pneumonia ENTITY .

Sinus rhythm ENTITY Borderline low QRS voltage ENTITY rsr pattern in leads V1 and V2 normal variant ENTITY Right precordialanterior T wave ENTITY changes are nonspecific ENTITY but clinical ENTITY correlation ENTITY is suggested Since previous tracing of, right precordial T wave ENTITY changes ENTITY more prominent

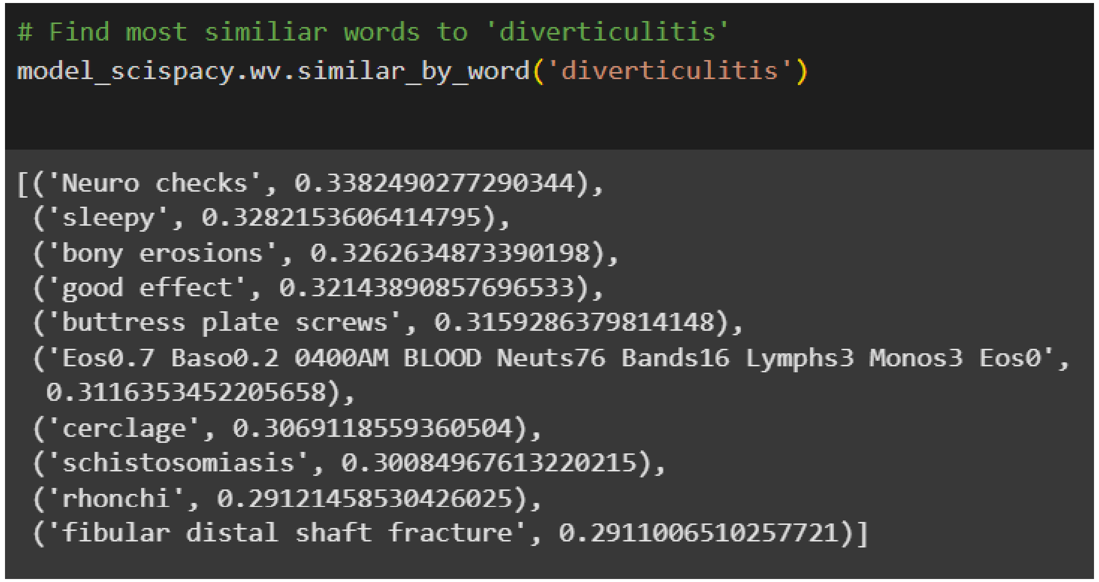
----

SCISPACY WORD2VEC

This word2vec by Scispacy seems to make a bit more sense than Spacy’s however it could still improve.

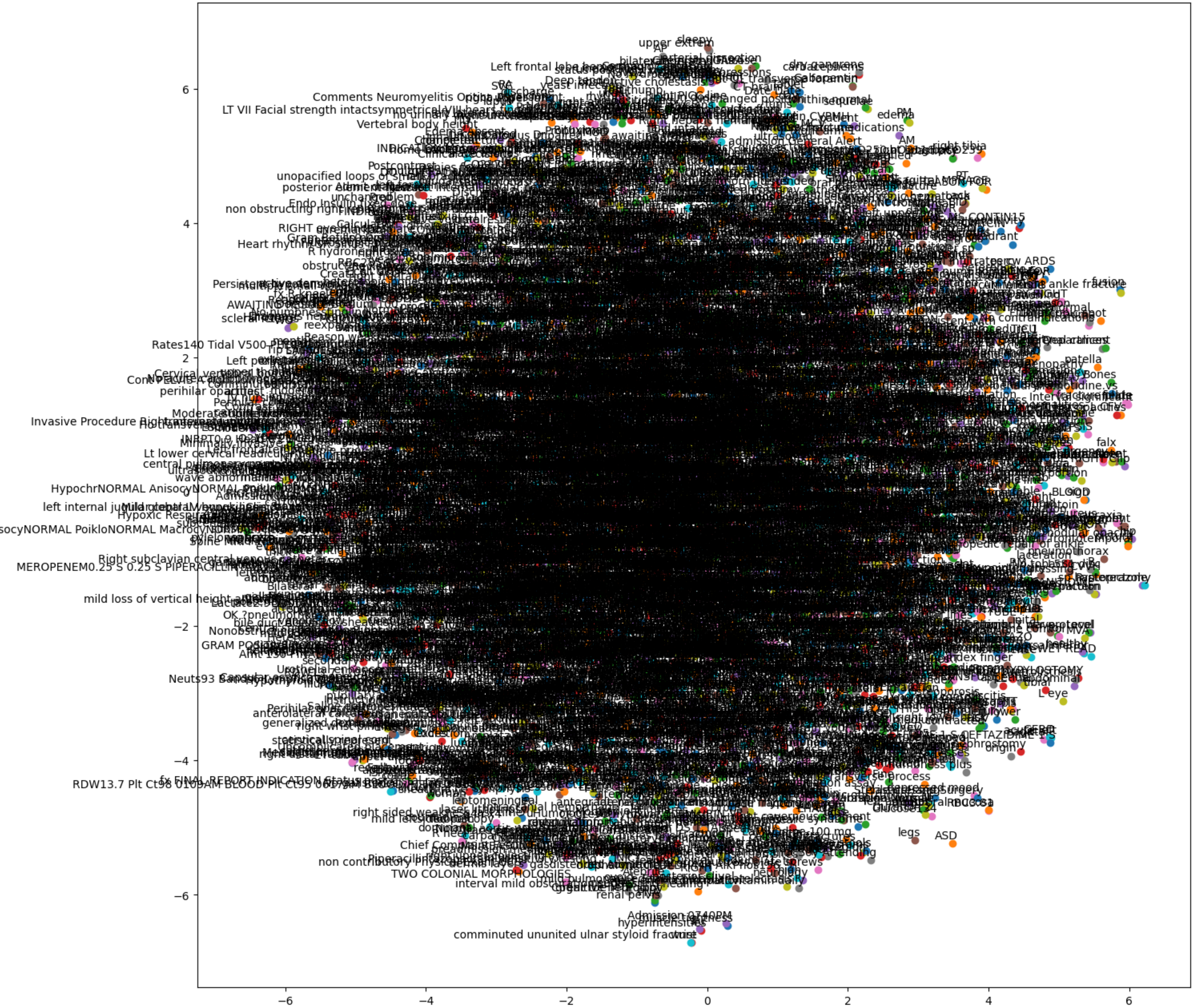


```
-7.0611173e-03, -1.5227579e-03,  2.6500411e-03,  3.8315416e-03,
 4.6273521e-03, -6.4738663e-03, -6.6725356e-03, -8.6745219e-03,
-9.5049599e-03,  4.6640079e-04,  4.9921954e-03,  5.1155253e-03,
-8.6483825e-03, -4.1282843e-03,  4.5276475e-03,  7.9027005e-03,
 5.1184972e-03, -1.3470954e-03, -3.5143544e-03, -9.2019998e-03,
 5.6368136e-03,  5.8939913e-03,  8.9649195e-03,  5.9732059e-03,
-1.0734015e-03, -7.5223846e-03,  4.5159226e-03, -8.3971135e-03,
 3.8024560e-03,  3.5979334e-05, -3.9827162e-03,  3.1726509e-03,
 5.2741994e-03,  9.6654259e-03, -6.7856698e-03,  2.1551482e-03,
-3.1155529e-03, -2.0836447e-03,  1.4447230e-03,  3.4121235e-03,
-1.0069066e-03,  5.7823109e-03,  7.4876416e-03,  1.6989755e-03,
-9.5217982e-03, -3.5139048e-03, -1.0162882e-03,  1.3274883e-03,
 4.6040667e-03, -1.6323101e-03, -5.6531169e-03,  9.6453552e-04,
 7.2515770e-03, -5.8803214e-03,  2.4550057e-03,  8.5401367e-03,
 8.7210555e-03,  2.9717474e-03, -8.0181994e-03,  4.8342687e-03,
-5.8901060e-04,  6.9626104e-03,  3.1928925e-03, -3.7182108e-03,
 7.4085034e-03, -9.3792444e-03, -4.0415670e-03, -8.6119082e-03,
```



```
# Find most similiar words to 'diverticulitis'
model_scispacy.wv.similar_by_word('diverticulitis')
```

```
[('Neuro checks', 0.3382490277290344),
 ('sleepy', 0.3282153606414795),
 ('bony erosions', 0.3262634873390198),
 ('good effect', 0.32143890857696533),
 ('buttress plate screws', 0.3159286379814148),
 ('Eos0.7 Baso0.2 0400AM BLOOD Neuts76 Bands16 Lymphs3 Monos3 Eos0',
  0.3116353452205658),
 ('cerclage', 0.3069118559360504),
 ('schistosomiasis', 0.30084967613220215),
 ('rhonchi', 0.29121458530426025),
 ('fibular distal shaft fracture', 0.2911006510257721)]
```

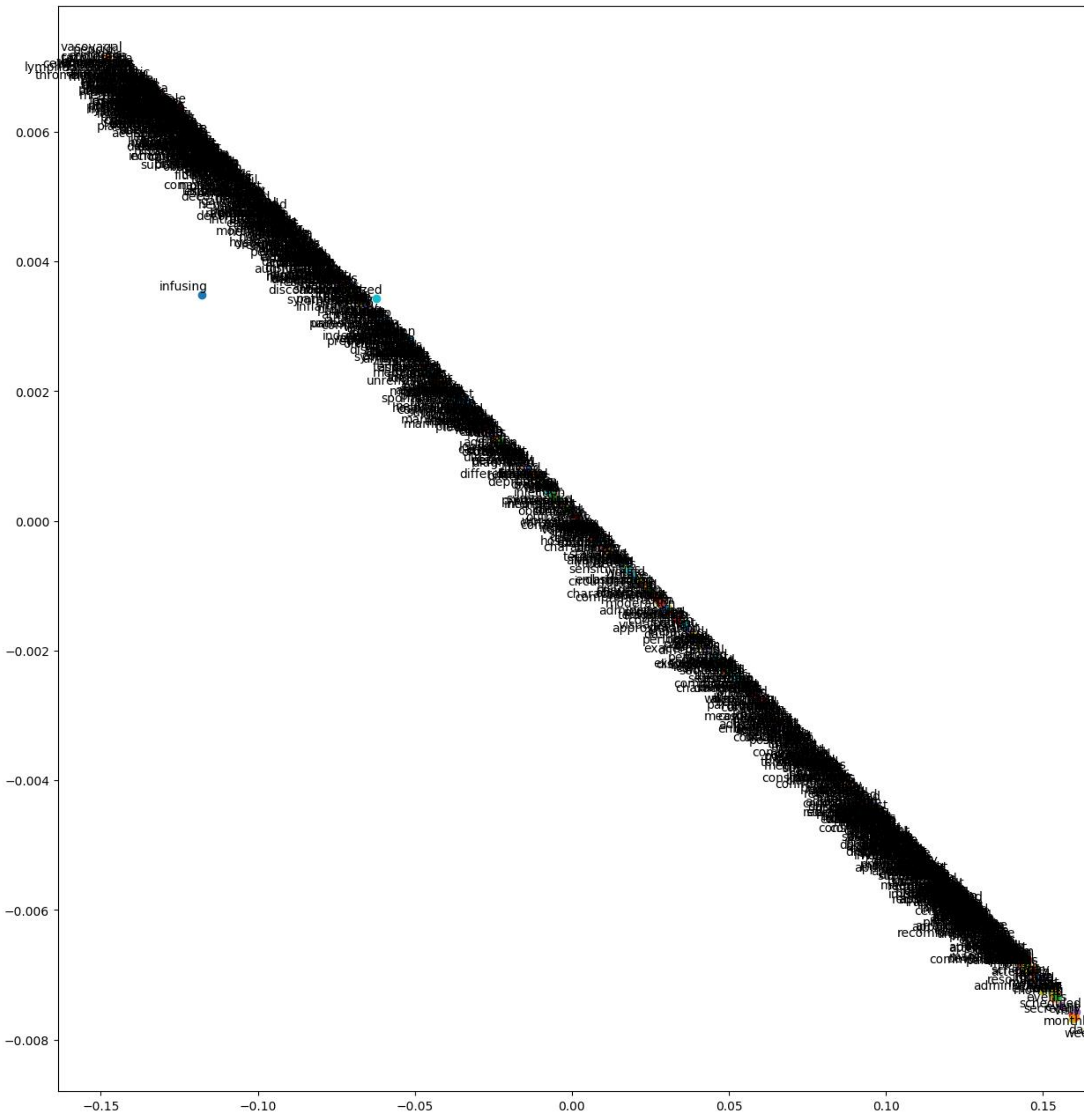


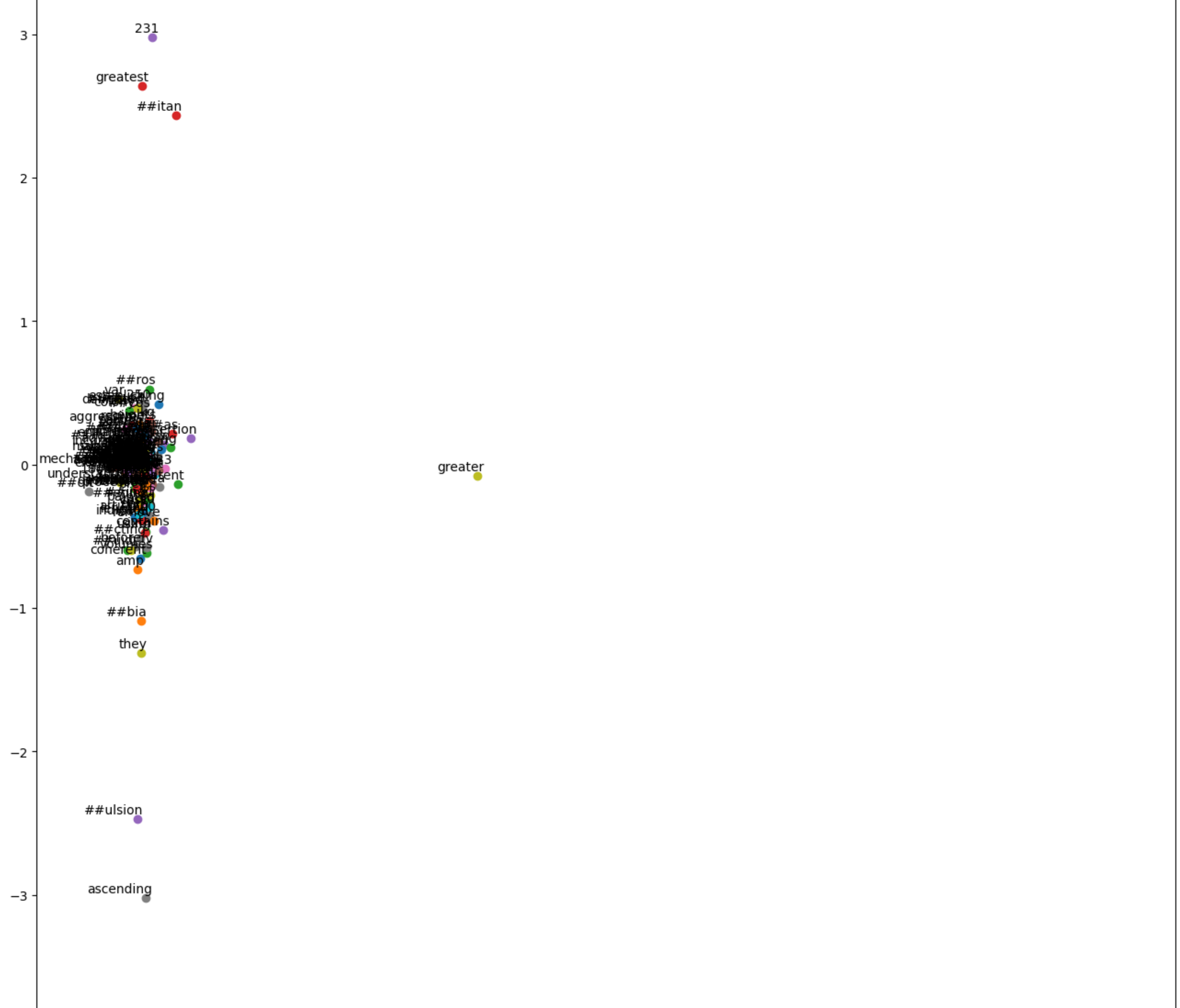
SISPACY TSNE PLOT

We can see from this that Scispacy was able to pick out many more entities that seem more relevant to the medical field. However, it is hard to confirm how well it does in mapping similarity between them.

SCISPACY TSNE PLOT WITH PRE-TRAINED MODEL

We can see again that more relevant entities were picked out but is too thick to evaluate. This also shows more words compared to mainly numbers.



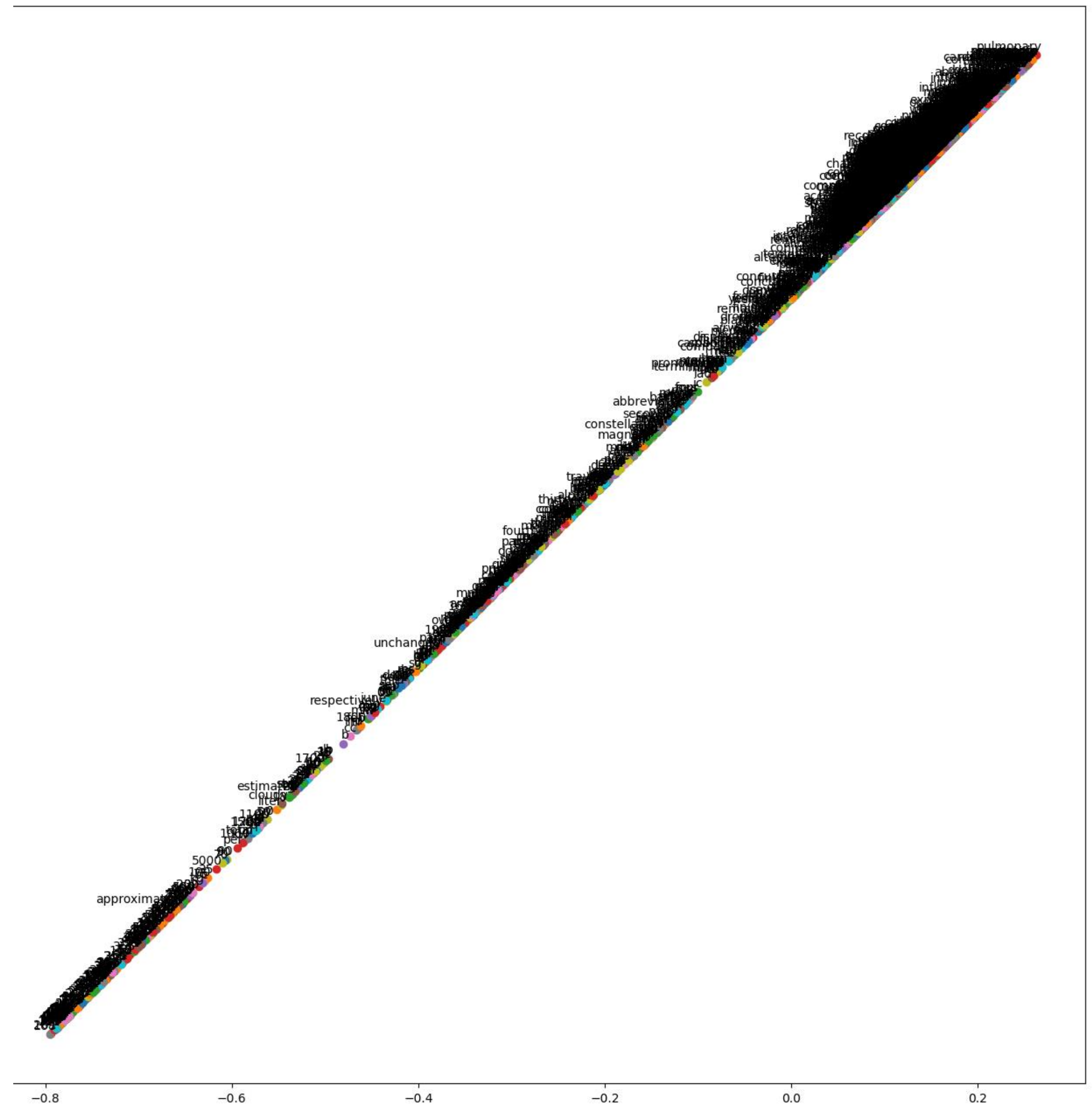


BLUEBERT TSNE PLOT

blueBert seemed to cluster more words more tightly and not include many entities.

BLUEBERT TSNE PLOT WITH PRE-TRAINED MODEL

For the pre-trained model, blueBert does fairly well separating most of the words and the numbers. Words like estimate and approximate are close to numbers which makes sense.



CONCLUSION

- Scispacy and blueBERT outperform Spacy in medical entity recognition.
- Scispacy captures more relevant terms but struggles with similarity mapping.
- blueBERT achieves better clustering but identifies fewer entities overall.
- Future improvements include fine-tuning word embeddings on MIMIC-III data and leveraging external biomedical knowledge sources like UMLS and SNOMED-CT.