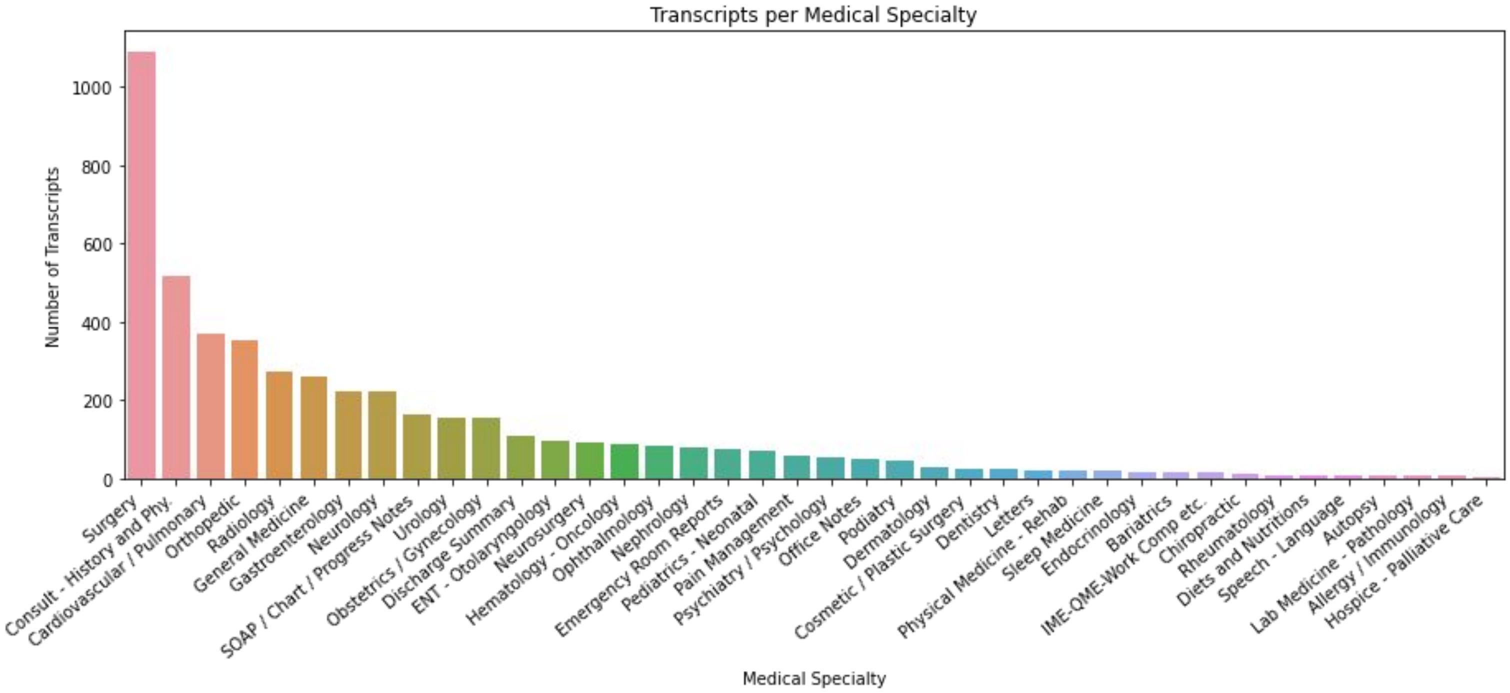
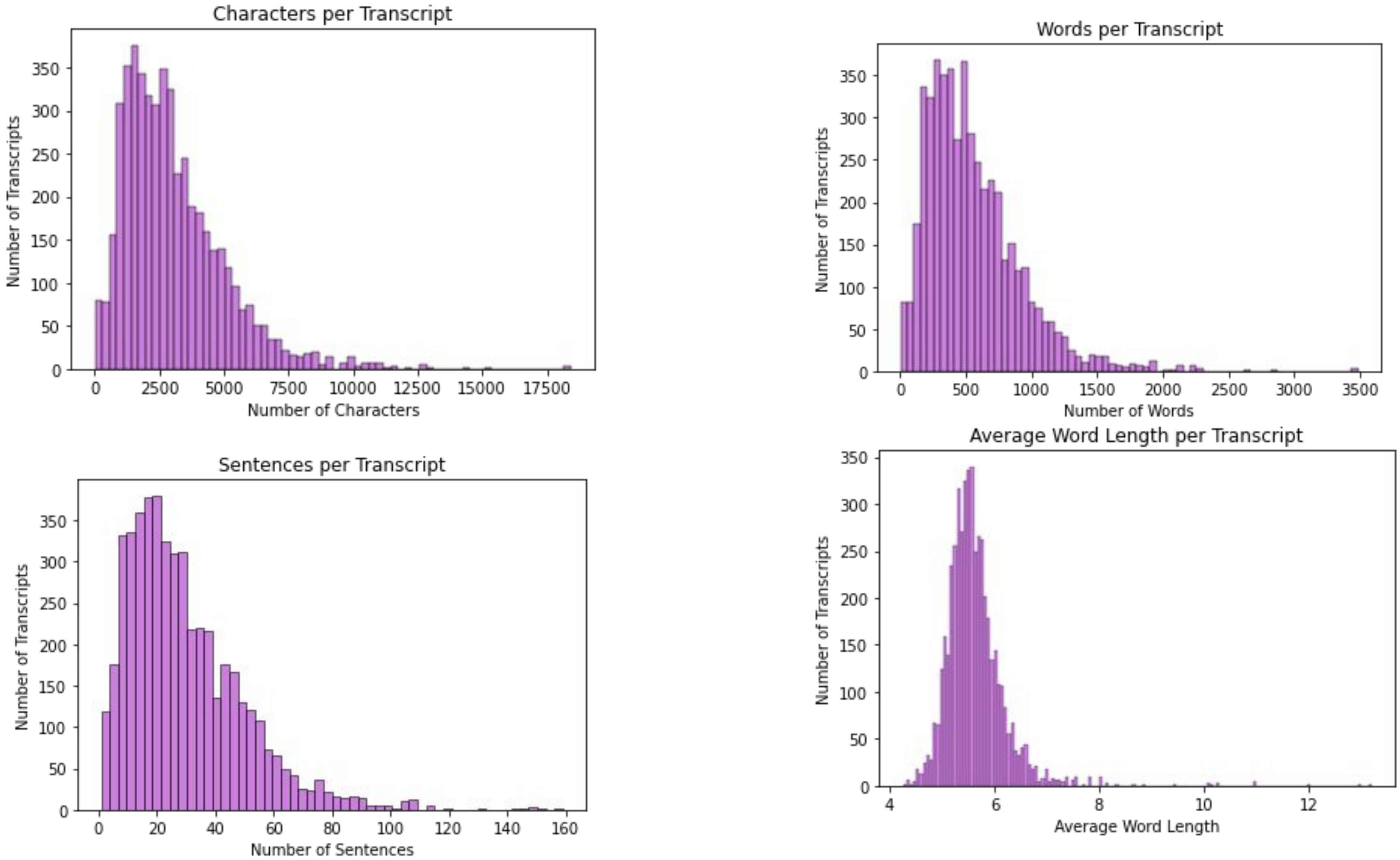
Medical Transcription Classification

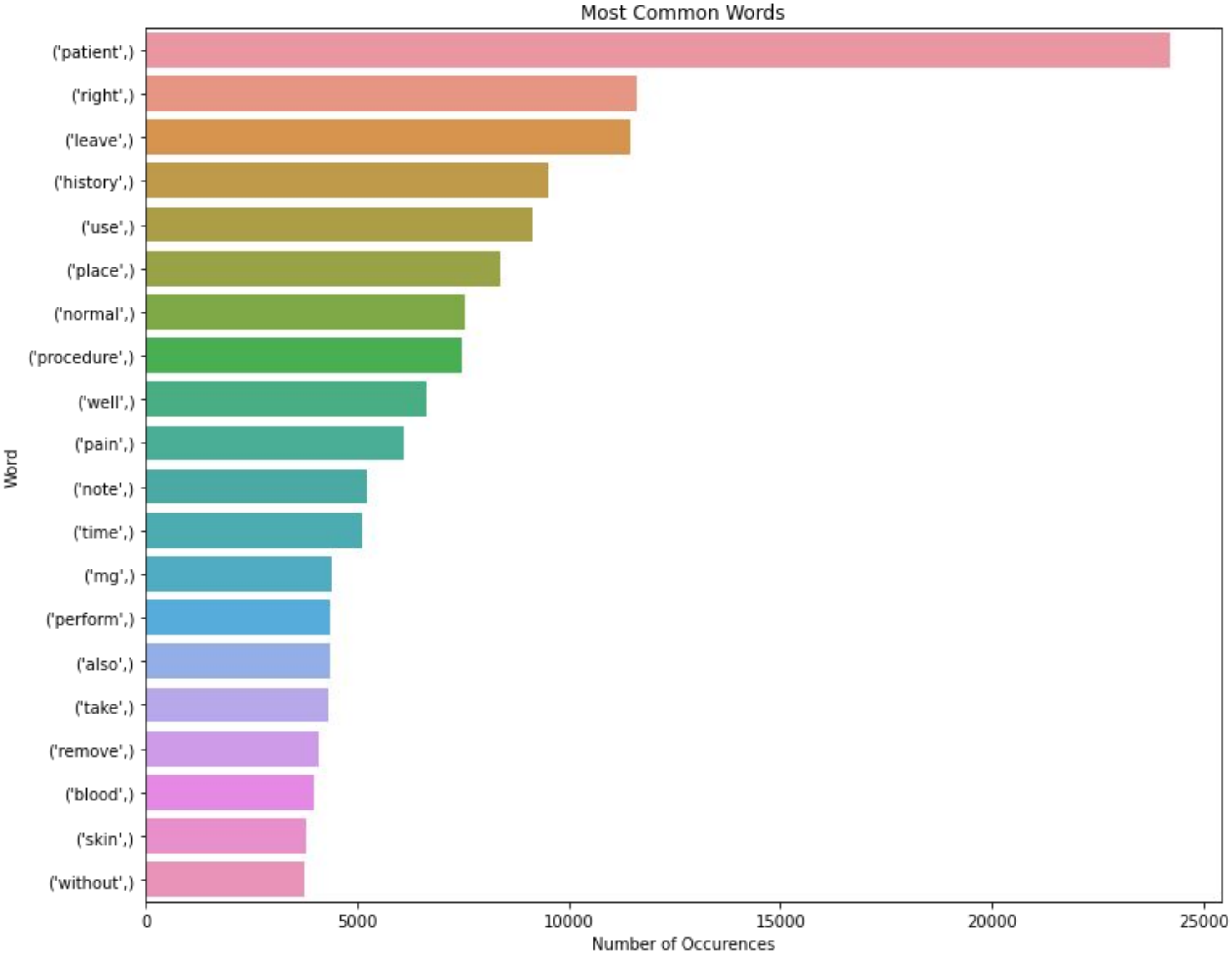
Freya Gray
CS39AA

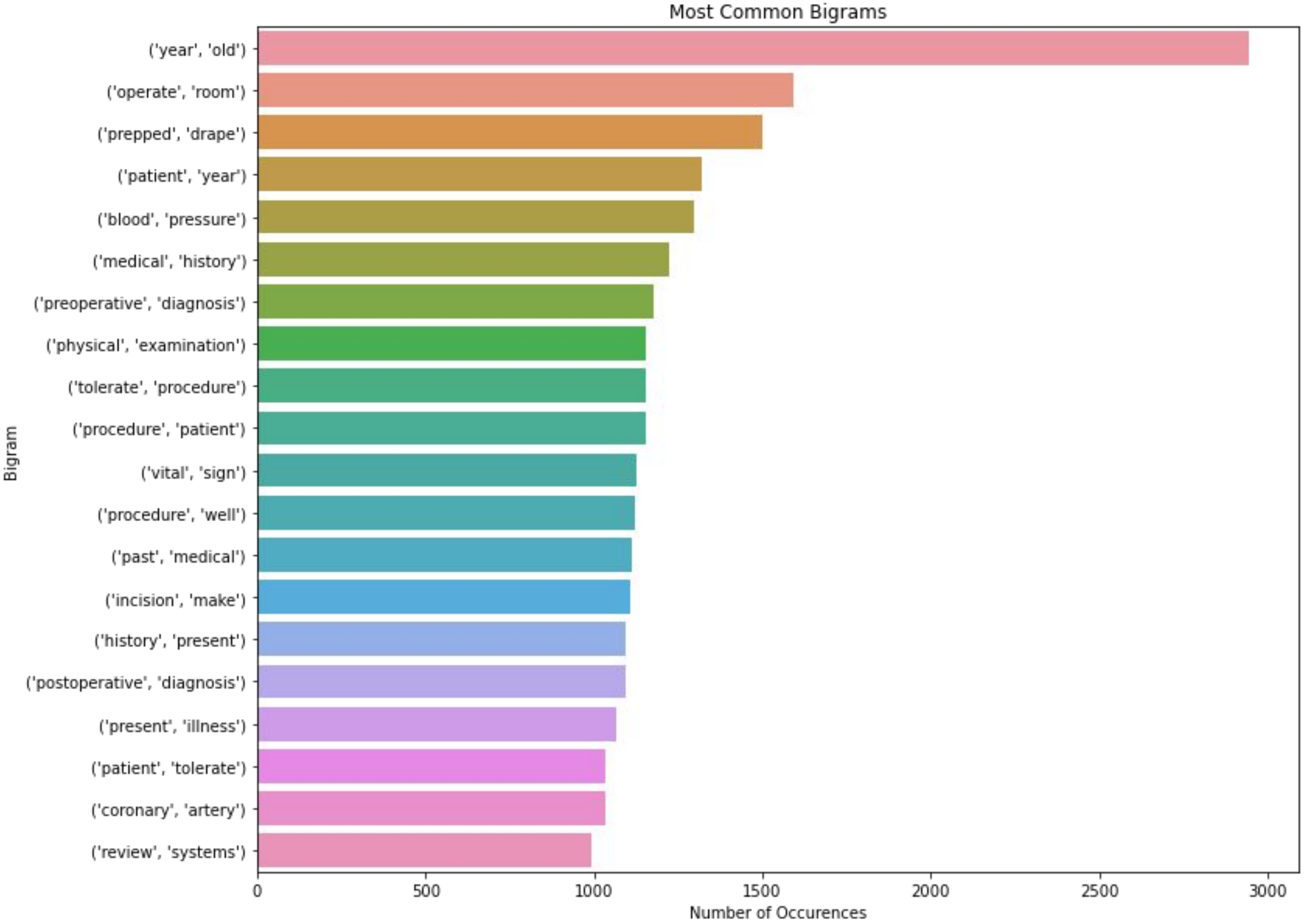
Dataset

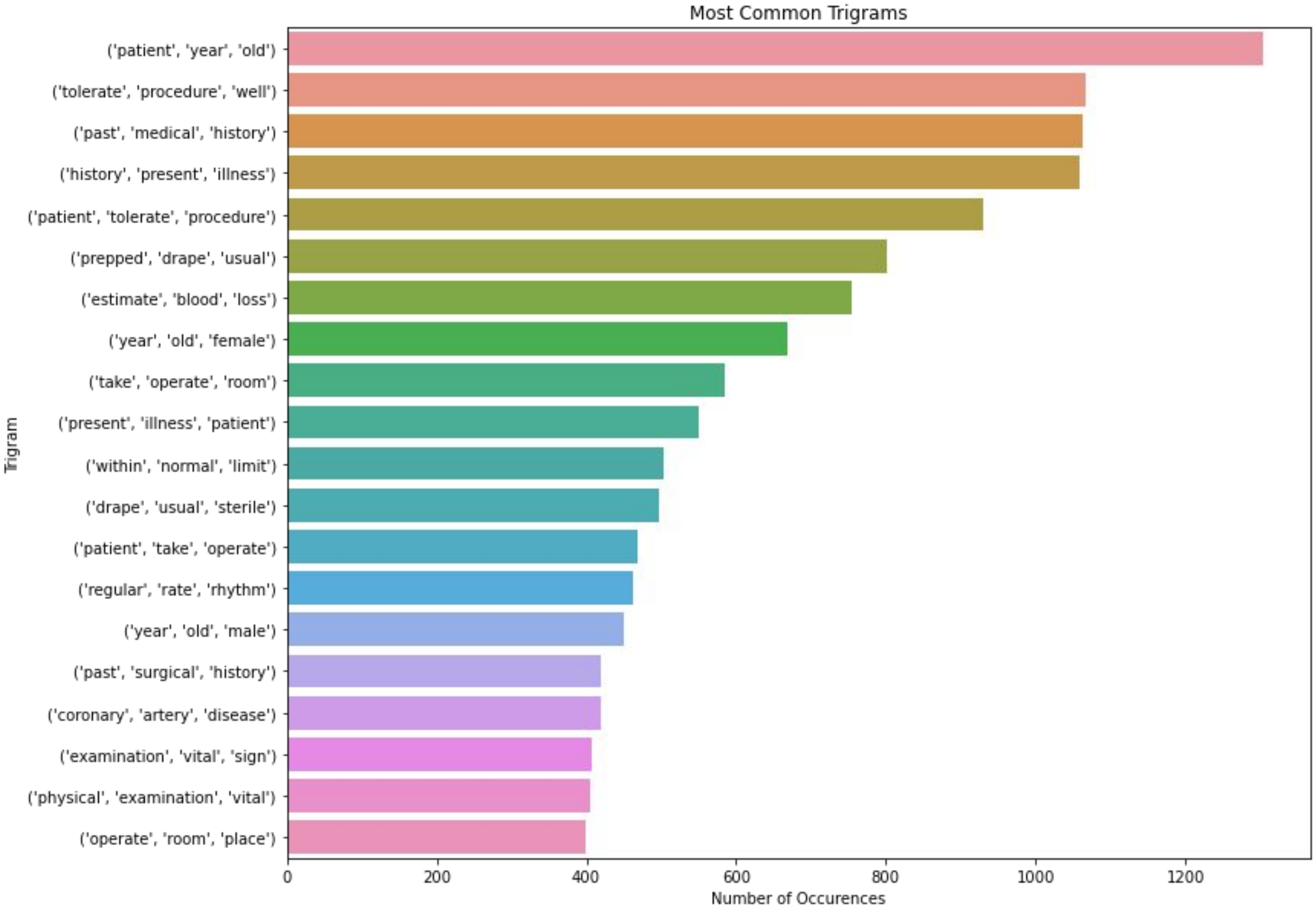
- Medical transcripts from mtsamples.com
- Transcripts from medical professionals after seeing a patient
- All names and dates have been changed or removed
- 40 different medical specialties
- 4966 transcripts
- Transcripts come in a wide range of formats and styles


Transcripts per Medical Specialty
Number of Transcripts
0
200
400
600
800
1000
Surgery
Consult - History and Phy.
Cardiovascular / Pulmonary
Orthopedic
Radiology
General Medicine
Gastroenterology
Neurology
SOAP / Chart / Progress Notes
Urology
Obstetrics / Gynecology
Discharge Summary
ENT - Otolaryngology
Neurosurgery
Hematology - Oncology
Ophthalmology
Nephrology
Emergency Room Reports
Pediatrics - Neonatal
Pain Management
Psychiatry / Psychology
Office Notes
Podiatry
Dermatology
Cosmetic / Plastic Surgery
Dentistry
Letters
Physical Medicine - Rehab
Sleep Medicine
Endocrinology
Bariatrics
IME-QME-Work Comp etc.
Chiropractic
Rheumatology
Diets and Nutritions
Speech - Language
Autopsy
Lab Medicine - Pathology
Allergy / Immunology
Hospice - Palliative Care
Medical Specialty


Characters per Transcript
Number of Transcripts
0
50
100
150
200
250
300
350
0
2500
5000
7500
10000
12500
15000
17500
Number of Characters
Words per Transcript
Number of Transcripts
0
50
100
150
200
250
300
350
0
500
1000
1500
2000
2500
3000
3500
Number of Words
Sentences per Transcript
Number of Transcripts
0
50
100
150
200
250
300
350
0
20
40
60
80
100
120
140
160
Number of Sentences
Average Word Length per Transcript
Number of Transcripts
0
50
100
150
200
250
300
350
4
6
8
10
12
Average Word Length

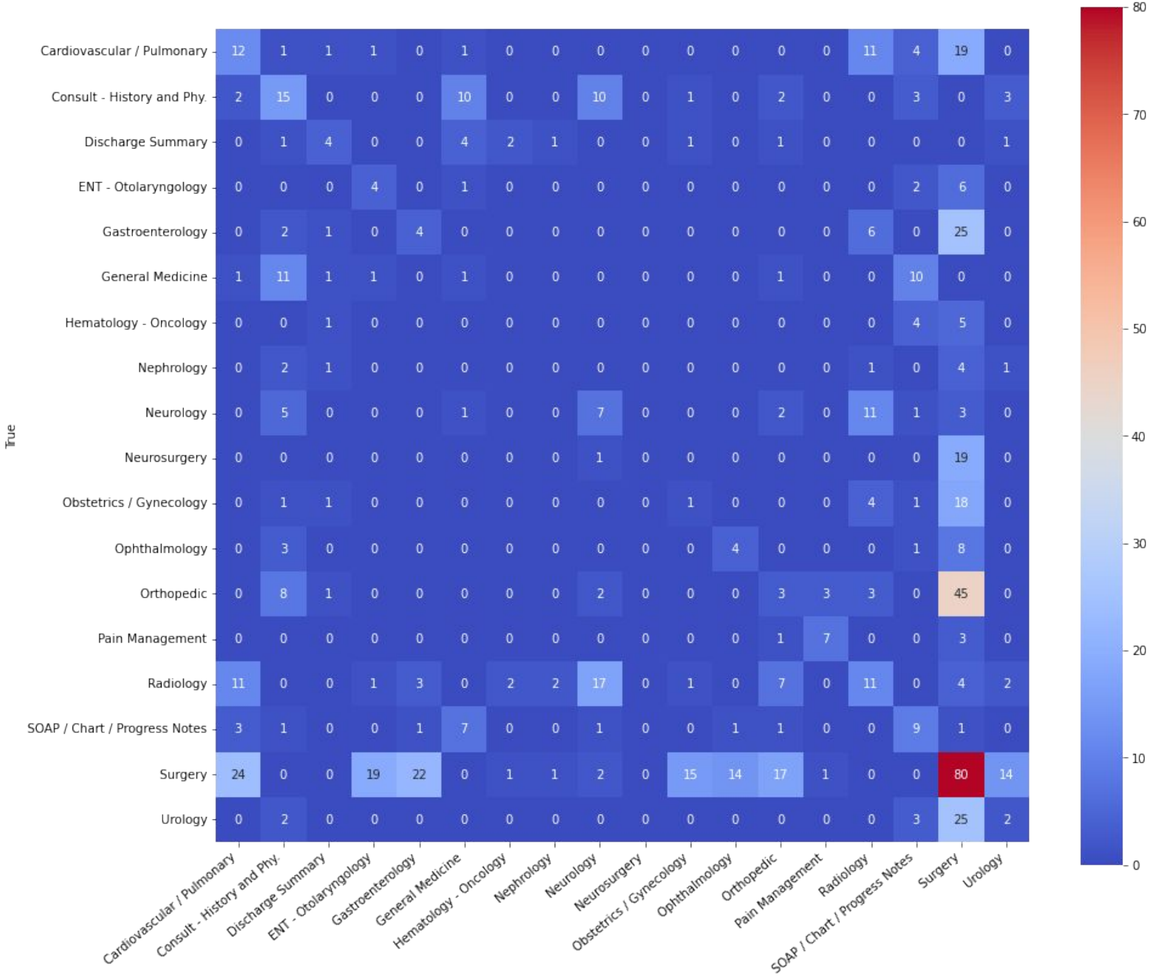

Most Common Words
('patient',)
('right',)
('leave',)
('history',)
('use',)
('place',)
('normal',)
('procedure',)
('well',)
('pain',)
('note',)
('time',)
('mg',)
('perform',)
('also',)
('take',)
('remove',)
('blood',)
('skin',)
('without',)
Word
0
5000
10000
15000
20000
25000
Number of Occurences


Most Common Bigrams
('year', 'old')
('operate', 'room')
('prepped', 'drape')
('patient', 'year')
('blood', 'pressure')
('medical', 'history')
('preoperative', 'diagnosis')
('physical', 'examination')
('tolerate', 'procedure')
('procedure', 'patient')
('vital', 'sign')
('procedure', 'well')
('past', 'medical')
('incision', 'make')
('history', 'present')
('postoperative', 'diagnosis')
('present', 'illness')
('patient', 'tolerate')
('coronary', 'artery')
('review', 'systems')
Bigram
0
500
1000
1500
2000
2500
3000
Number of Occurences


Most Common Trigrams
Trigram
('patient', 'year', 'old')
('tolerate', 'procedure', 'well')
('past', 'medical', 'history')
('history', 'present', 'illness')
('patient', 'tolerate', 'procedure')
('prepped', 'drape', 'usual')
('estimate', 'blood', 'loss')
('year', 'old', 'female')
('take', 'operate', 'room')
('present', 'illness', 'patient')
('within', 'normal', 'limit')
('drape', 'usual', 'sterile')
('patient', 'take', 'operate')
('regular', 'rate', 'rhythm')
('year', 'old', 'male')
('past', 'surgical', 'history')
('coronary', 'artery', 'disease')
('examination', 'vital', 'sign')
('physical', 'examination', 'vital')
('operate', 'room', 'place')
0
200
400
600
800
1000
1200
Number of Occurences

Baseline Model

- Reduced medical specialties to 18
- Support Vector Machine
- GridSearchCV to find best parameters
    - Best parameters: C = 1, kernel = linear
- Accuracy 23%

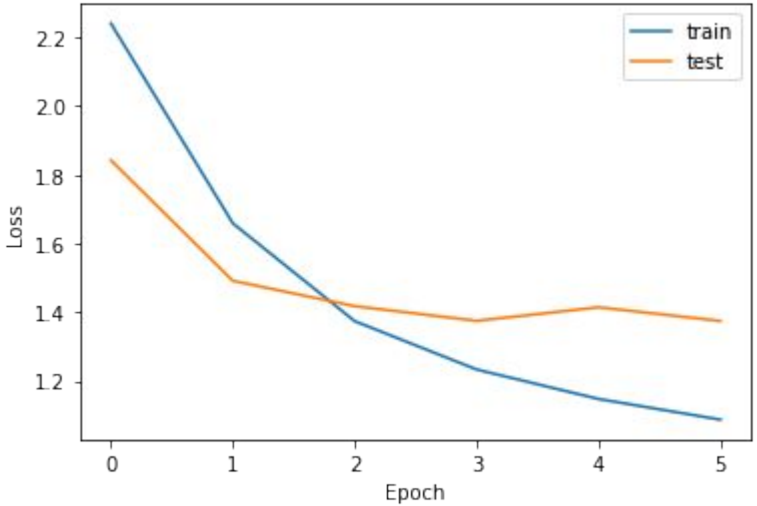


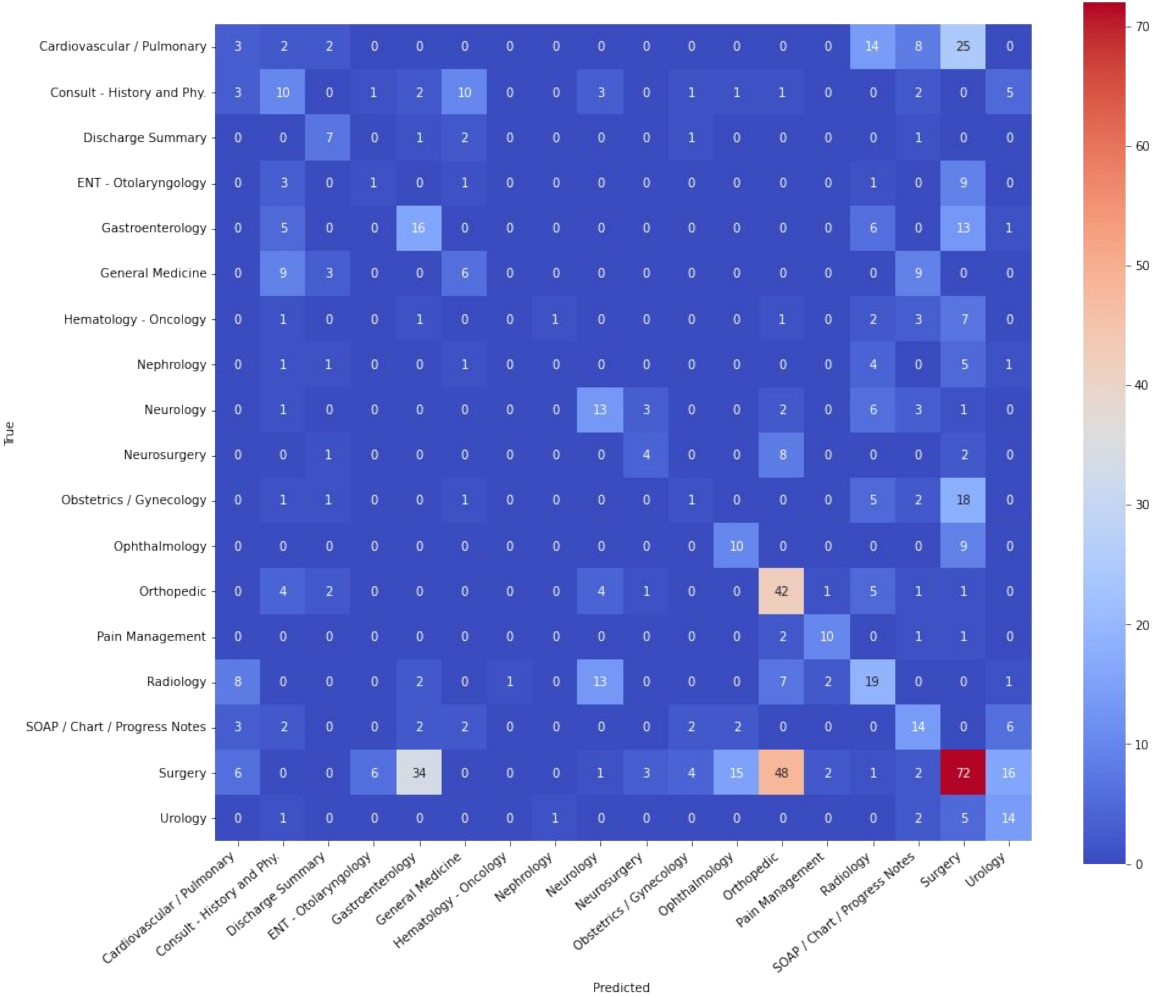
	Cardiovascular / Pulmonary	Consult - History and Phy.	Discharge Summary	ENT - Otolaryngology	Gastroenterology	General Medicine	Hematology - Oncology	Nephrology	Neurology	Neurosurgery	Obstetrics / Gynecology	Ophthalmology	Orthopedic	Pain Management	Radiology	SOAP / Chart / Progress Notes	Surgery	Urology
Cardiovascular / Pulmonary	12	1	1	1	0	1	0	0	0	0	0	0	0	0	11	4	19	0
Consult - History and Phy.	2	15	0	0	0	10	0	0	10	0	1	0	2	0	0	3	0	3
Discharge Summary	0	1	4	0	0	4	2	1	0	0	1	0	1	0	0	0	0	1
ENT - Otolaryngology	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	2	6	0
Gastroenterology	0	2	1	0	4	0	0	0	0	0	0	0	0	0	6	0	25	0
General Medicine	1	11	1	1	0	1	0	0	0	0	0	0	1	0	0	10	0	0
Hematology - Oncology	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4	5	0
Nephrology	0	2	1	0	0	0	0	0	0	0	0	0	0	0	1	0	4	1
Neurology	0	5	0	0	0	1	0	0	7	0	0	0	2	0	11	1	3	0
Neurosurgery	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	19	0
Obstetrics / Gynecology	0	1	1	0	0	0	0	0	0	0	1	0	0	0	4	1	18	0
Ophthalmology	0	3	0	0	0	0	0	0	0	0	0	4	0	0	0	1	8	0
Orthopedic	0	8	1	0	0	0	0	0	2	0	0	0	3	3	3	0	45	0
Pain Management	0	0	0	0	0	0	0	0	0	0	0	0	1	7	0	0	3	0
Radiology	11	0	0	1	3	0	2	2	17	0	1	0	7	0	11	0	4	2
SOAP / Chart / Progress Notes	3	1	0	0	1	7	0	0	1	0	0	1	1	0	0	9	1	0
Surgery	24	0	0	19	22	0	1	1	2	0	15	14	17	1	0	0	80	14
Urology	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	25	2

Accuracy:0.2332859174964438
Precision:0.20922610493758473
Recall:0.2332859174964438
F1 score:0.21742526566778728

RoBERTa

- Reduced medical specialties to 18
- Model limited max sequence length
- Just transcriptions
- Accuracy: 48.06%
- Loss:1.0876
- Validation Accuracy: 34.42%
- Validation Loss: 1.3746

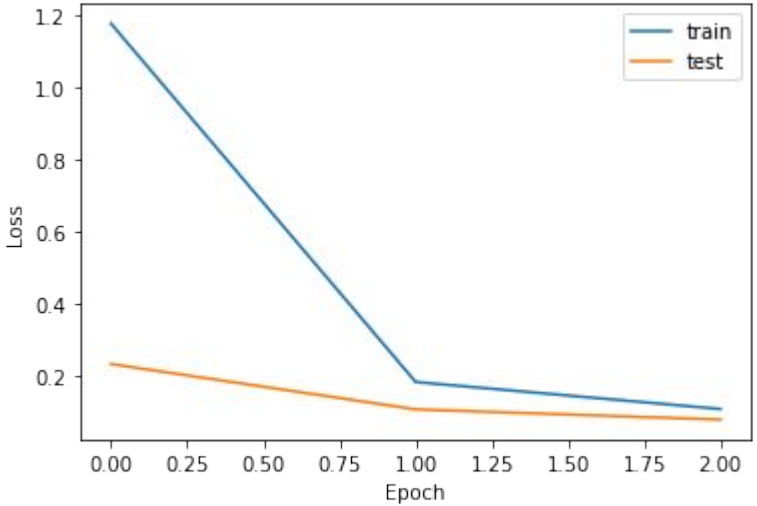


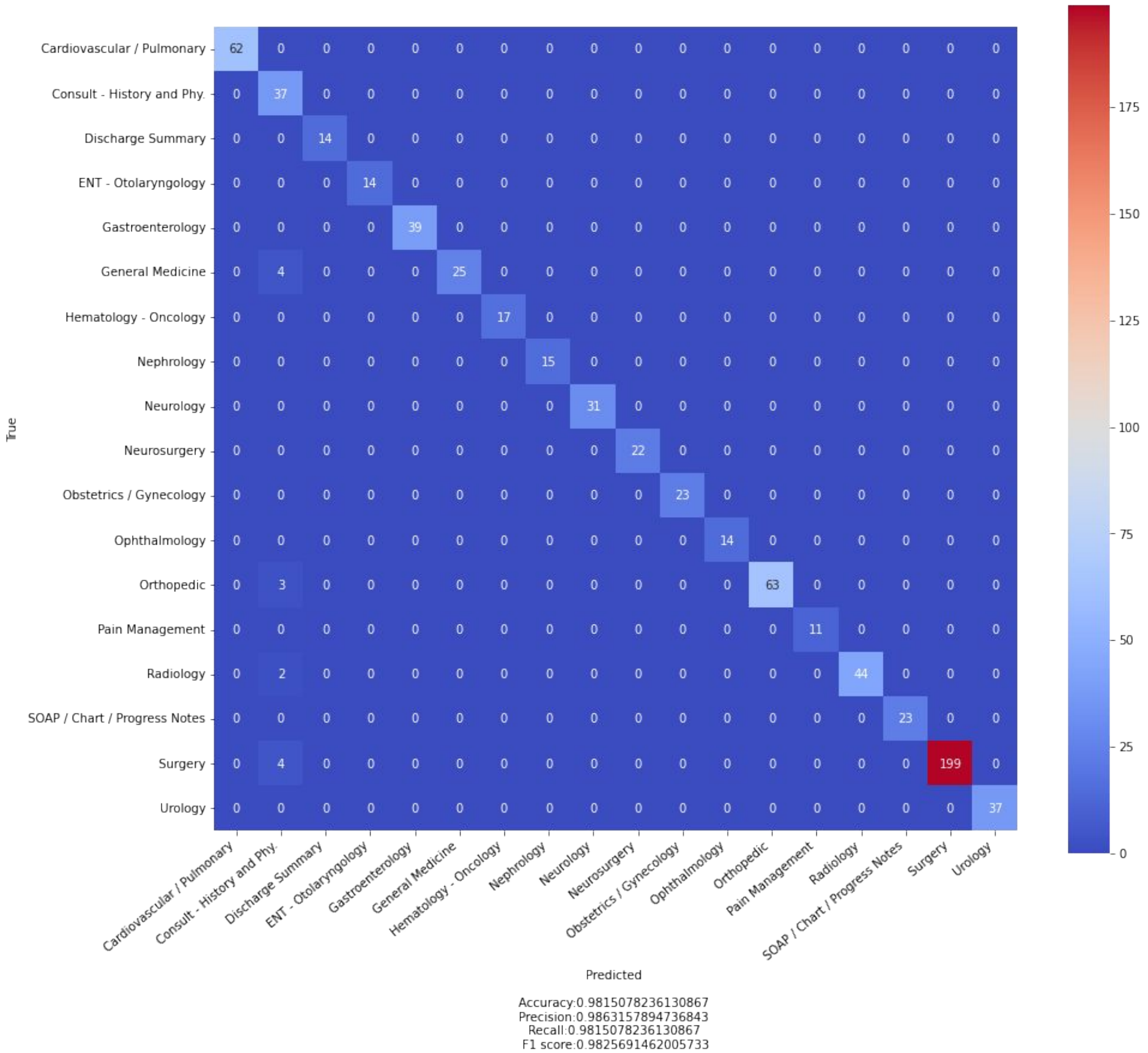


Accuracy:0.3442389758179232
Precision:0.31842947125638177
Recall:0.3442389758179232
F1 score:0.3199623527339034

RoBERTa

- Keywords
- Accuracy: 98.29%
- Loss: .1126
- Validation Accuracy: 98.58%
- Validation Loss: .0667





True \ Predicted	Cardiovascular / Pulmonary	Consult - History and Phy.	Discharge Summary	ENT - Otolaryngology	Gastroenterology	General Medicine	Hematology - Oncology	Nephrology	Neurology	Neurosurgery	Obstetrics / Gynecology	Ophthalmology	Orthopedic	Pain Management	Radiology	SOAP / Chart / Progress Notes	Surgery	Urology
Cardiovascular / Pulmonary	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Consult - History and Phy.	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Discharge Summary	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENT - Otolaryngology	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gastroenterology	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0
General Medicine	0	4	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0
Hematology - Oncology	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0
Nephrology	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0
Neurology	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0
Neurosurgery	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0
Obstetrics / Gynecology	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0
Ophthalmology	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0
Orthopedic	0	3	0	0	0	0	0	0	0	0	0	0	63	0	0	0	0	0
Pain Management	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0
Radiology	0	2	0	0	0	0	0	0	0	0	0	0	0	0	44	0	0	0
SOAP / Chart / Progress Notes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0
Surgery	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	199	0
Urology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37

Accuracy:0.9815078236130867
Precision:0.9863157894736843
Recall:0.9815078236130867
F1 score:0.9825691462005733

Conclusions

- Pre-trained model outperformed the baseline model
- Overlap between specialties seems to have increased errors
- Possible Improvements
    - Different Model
    - Better data cleaning
    - Classify transcripts differently

Sources

- Dataset
- Hugging Face Guide
- Text Cleaning
- Data Exploration: link 1, link 2