MedTextClassifier - Knowledge based Word Sense Disambiguation for classifying clinical text

Sahil Aneja | Rahul Ananda Bijai
Department of Computer Science,
St. Francis Xavier University, Nova Scotia, Canada

Abstract

There has been a rapid adoption of EHR (Electronic Health Records) across many medical institutions and a huge amount of EHR data is lying unused. The EHRs such as doctor's prescription, clinical notes and discharge notes are usually unstructured and difficult to process. A good machine learning model generally demands a significant amount of human labour to generate labelled training data and execute feature engineering. Hence, we made use of NLP techniques such as Word Sense Disambiguation to extract useful information from this huge raw text data. We applied knowledge-based techniques for WSD that make use of data that already exist in the form of dictionaries or sense inventories such as UMLS (Unified Medical Language System). For performing WSD, we understood and used the Apache cTAKES tool and hence we were able to extract several medical concepts from the unstructured raw text. These medical concepts were then used to generate the training dataset. On checking our results using several models, we noticed that Logistics Regression was able to predict the 'Medical Speciality' based on the medical transcription with a good accuracy. Here, we studied how knowledge-based WSD when combined with other machine learning algorithms can help to perform multi-class classification of clinical documents.

Introduction

Ambiguity is a big challenge while processing the clinical data for analysis. In natural language, there could be a lot of ambiguous terms i.e. terms having more than one meaning. For example, the word 'discharge' could mean 'a point at which the patient leaves the hospital' or 'a body fluid'. Similarly, the word 'surgery' could mean 'a branch of medicine that applies operative procedures' or 'the operative procedure itself'. Identifying the true meaning of such ambiguous terms is the main challenge while extracting information from biomedical documents.

Because English contains ambiguous concepts that might be difficult to decipher, Word Sense Disambiguation (WSD) is a significant difficulty for any computerized text processing system. It is important to develop natural language processing (NLP) methods that can accurately analyze this data in a reasonable amount of time to gain insights from it. Many high-level information extraction and knowledge discovery applications rely on NLP components such as named entity recognition programs, syntactic parsers, and relation extractors.

WSD is an open problem in the area of NLP and further breakthroughs in this area will be helpful to many applications based on NLP. There are several approaches to perform WSD - Supervised, Semi-supervised and Knowledge based. Supervised and Semi-supervised approaches require a labelled training dataset beforehand. Since, we are directly working on raw clinical text, we follow the Knowledge based approach that utilizes UMLS to generate medical concepts.

Document classification is one of the main applications of WSD in the biomedical domain. In this research, we study the medical transcription data and extract useful concepts that can further help us to classify the clinical data. Here, we will try to predict the medical specialty based on the patient's data so that the patient can be referred to the right department/doctor.

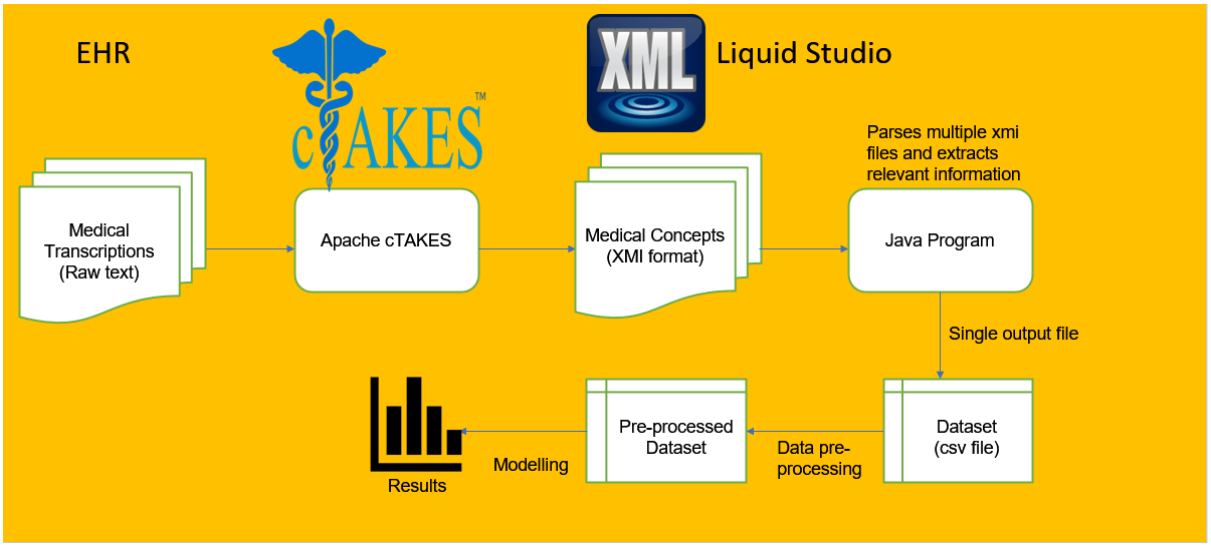
Materials

We explored NLP software tools such as MetaMap and Apache cTAKES to generate biomedical concept entities from the clinical text. A common feature of these tools is the NER (Named Entity Recognition) task that is specific to the biomedical domain. Both cTAKES and MetaMap derive the biomedical concepts from the UMLS. UMLS is a large repository of biomedical vocabularies that gets updated every year. We found that results from cTAKES were more interpretable in a biomedical context. So, we decided to choose cTAKES over MetaMap and used it's results for further data processing. In order to use cTAKES, it is mandatory to have a UMLS license.

Methodology

Let's have a look at all the methods used in this study involving Knowledge based WSD and Machine Learning models.

Architecture



The series of steps involved are as follows:

- EHR documents (Medical Transcriptions from mtsamples) in raw text format are passed on to the Apache cTAKES tool.
- Apache cTAKES internally used the UMLS Metathesaurus to generate medical concepts. The output of this tool is in the form of xmi files. For each input .txt file, a corresponding .xmi file is generated.
- Developed a Java program that parses all the xmi files, extracts the relevant information such as Diseases, Medications, Symptoms and Procedures and generates a single output csv file.
- This csv file is then further pre-processed to generate the final training dataset.
- Machine Learning models are then applied on the dataset to perform multi-class classification and generate the results.

Dataset Preparation and Pre-processing

Step1: Process raw text using cTAKES

The initial data is the free-text clinical notes/medical transcriptions. For our study we picked around 5000 medical transcriptions from the mtsamples dataset. A sample medical transcription(**Input**) is shown below:

```
SUBJECTIVE:,  This 23-year-old white female presents with complaint of allergies.
She used to have allergies when she lived in Seattle but she thinks they are worse here.
In the past, she has tried Claritin, and Zyrtec.  Both worked for short time but then
seemed to lose effectiveness.  She has used Allegra also.  She used that last summer
and she began using it again two weeks ago.  It does not appear to be working very well.
She has used over-the-counter sprays but no prescription nasal sprays.  She does have
asthma but doest not require daily medication for this and does not think it is flaring
up.,MEDICATIONS: , Her only medication currently is Ortho Tri-Cyclen and the Allegra.,
ALLERGIES: , She has no known medicine allergies.,OBJECTIVE:,Vitals:  Weight was 130
pounds and blood pressure 124/78.,HEENT:  Her throat was mildly erythematous without
exudate.  Nasal mucosa was erythematous and swollen.  Only clear drainage was seen.
TMs were clear.,Neck:  Supple without adenopathy.,Lungs:  Clear.,ASSESSMENT:,
Allergic rhinitis.,PLAN:,1.  She will try Zyrtec instead of Allegra again.
Another option will be to use loratadine.  She does not think she has prescription
coverage so that might be cheaper.,2.  Samples of Nasonex two sprays in each nostril
given for three weeks.  A prescription was written as well.
```

The input to cTAKES will be a text file containing clinical data(as shown above) and the output will be in the form of an xmi(xml metadata interchange) file. This output xmi file will provide us with a list of annotations such as ‘DiseaseDisorderMention’, ‘MedicationMention’, ‘SignSymptomMention’ and ‘UMLSConcept’ which can be used further for the classification of biomedical data.

Procedure to use cTAKES

The prerequisite is to have a valid UMLS license.

Run the default clinical pipeline command:

bin/runClinicalPipeline -i <input-directory-name> --xmiOut <output-directory-name> --key <umls-api-key>

```
C:\apache-ctakes-4.0.0.1\bin>runClinicalPipeline --key 79574097-7d87-426f-8678-331244f0ee8e -i C:\AllTxtFiles
 --xmiOut C:\AllXmiFiles\
log4j: reset attribute= "false".
log4j: Threshold ="null".
log4j: Retreiving an instance of org.apache.log4j.Logger.
log4j: Setting [ProgressAppender] additivity to [false].
log4j: Level value for ProgressAppender is  [INFO].
log4j: ProgressAppender level set to INFO
log4j: Class name: [org.apache.log4j.ConsoleAppender]
log4j: Parsing layout of class: "org.apache.log4j.PatternLayout"
log4j: Setting property [conversionPattern] to [%m].
log4j: Adding appender named [noEolAppender] to category [ProgressAppender].
log4j: Retreiving an instance of org.apache.log4j.Logger.
```

In case we need to add a custom dictionary, we can run the piper file as follows:

bin\runPiperFile --key -i <input-directory-name> --xmiOut <output-directory-name> -p <path-piper-file> -l <path-dictionary-file>

```
.............................
Loading model:
.........................
24 Oct 2021 19:22:33  INFO SentenceDetector - Starting processing.
24 Oct 2021 19:22:33  INFO TokenizerAnnotatorPTB - process(JCas) in org.apache.ctakes.core.ae.TokenizerAnnotatorPTB
24 Oct 2021 19:22:33  INFO ContextDependentTokenizerAnnotator - process(JCas)
24 Oct 2021 19:22:33  INFO POSTagger - process(JCas)
24 Oct 2021 19:22:33  INFO Chunker -  process(JCas)
24 Oct 2021 19:22:35  INFO ChunkAdjuster -  process(JCas)
24 Oct 2021 19:22:35  INFO ChunkAdjuster -  process(JCas)
24 Oct 2021 19:22:35  INFO AbstractJCasTermAnnotator - Starting processing
24 Oct 2021 19:22:35  INFO AbstractJCasTermAnnotator - Finished processing
24 Oct 2021 19:22:35  INFO ClearNLPSemanticRoleLabelerAE - Starting processing
24 Oct 2021 19:22:35  INFO ClearNLPSemanticRoleLabelerAE - Finished processing
```

Here, the value of the ‘key’ attribute will be the API key corresponding to your UMLS account (API key can be found in the Profile section). It usually takes 3 days to get the account verified by the National Library of Medicine team.

Here is the view of the output xmi file.

<..> UmlsConcept (77)

	id	codingScheme	code	score	disam...	cui	tui	preferredText
1	4487	RXNORM	203576	0.0	false	C0701055	T121	Claritin
2	4477	RXNORM	203576	0.0	false	C0701055	T109	Claritin
3	4595	RXNORM	58930	0.0	false	C0162723	T121	Zyrtec

<..> DiseaseDisorderMention (4)

	id	sofa	begin	end	id	ontologyConceptArr	typeID	discov...	confid...	polarity
1	5097	1	520	526	0	5084	2	1	0.0	1
2	5053	1	1028	1045	0	5040	2	1	0.0	1
3	5141	1	1037	1045	0	5128	2	1	0.0	1

<..> MedicationMention (16)

	id	sofa	begin	end	id	ontologyConceptArr	typeID	discov...	confid...	polarity
1	4501	1	200	208	0	4487 4477	1	1	0.0	1
2	4609	1	214	220	0	4595 4585	1	1	0.0	1
3	4728	1	303	310	0	4704 4714	1	1	0.0	1

We can find whether a UMLS concept denotes a symptom, disease, procedure or medication by mapping the 'id' field in the UmlsConcept table with the 'ontologyConceptArr' in other tables. In the above xmi view we can see that UMLS concept with id '4595' denotes a Medication. Therefore, Zyrtec is one of the medications related to the processed transcription.

Step 2: JAVA Program - Parse the XMI files and Extract relevant information

We developed a Java project that takes XMI files as input, parses all the files, extracts relevant information(Transcription, Symptoms, Diseases, Procedures and Medications, Preferred Text(UMLS Name) and generates a single output file in csv format.

Output: single csv file

transcription	symptoms	diseases	procedures	medications
SUBJECTIVE:, This 23-year-old white fem	Blood Pressure, Lymphadenopathy,	Rhinitis, Infantile Neuroaxonal Dystrophy,	Weighing patient, Transcranial magnetic	stimulation, Drainage procedure
DESCRIPTION:,1. Normal cardiac chambe	effusion, Ejection as a Sports activity	Tricuspid Valve Insufficiency, Pericardial e	Doppler studies	
PREOPERATIVE DIAGNOSIS: , Right inguin	Pre-op diagnosis, Pain, Infertility, Dr	Retinitis Pigmentosa, Hernia sac, Hernia, I	Spinal Anesthesia, Dressing of skin or wc	Solution Dosage Form, Sponge Do
PREOPERATIVE DIAGNOSIS: , Possible infl	Able (finding), Pre-op diagnosis, Pos	Disease, Inflammatory Bowel Diseases, In	Sedation procedure, Interventional procedure, colonoscopy, Resection of p	
PROCEDURE: , Colonoscopy.,PREOPERATI	No family history of, No abdominal p	Dyslipidemias, Disease, Hypertensive dise	Screening colonoscopy, Screening for cancer, Interventional procedure, col	
PROCEDURE: , Colonoscopy.,PREOPERATI	Rectal hemorrhage, Perirectal absce	Posterior pituitary disease, Perianal absce	Interventional procedure, colonoscopy,	Rectal Dosage Form, Antibiotics
PREOPERATIVE DIAGNOSIS: , Colon polyp	Rest, Pre-op diagnosis, Actual blood	Colonic Polyps, Hemorrhoids, Internal her	Digital Rectal Examination, Interventiona	Rectal Dosage Form
PREOPERATIVE DIAGNOSIS: , History of p	Blood Pressure, Morphology within	Diverticular disease of colon, Disease, Pro	Electrocardiographic monitoring, Interventional procedure, colonoscopy, T	
PROCEDURE IN DETAIL: , Following instru	Vision, Skin appearance normal (finding), Gait normal		Interventional procedure, colonoscopy,	Oral Dosage Form
PREOPERATIVE DIAGNOSIS: , Screening. ,F	Pre-op diagnosis, Post-op diagnosis,	bite injury, Infantile Neuroaxonal Dystrop	Digital Rectal Examination, Screening for	Rectal Dosage Form
PREOPERATIVE DIAGNOSES:, Prior histor	H/O: anemia, Abdominal bloating, S	Anemia, Hemorrhoids, External hemorrh	Digital Rectal Examination, Interventiona	Rectal Dosage Form
MEDICATIONS:,1. Versed intravenously.,	Skin appearance normal (finding),	Complication, Gait normal, Sedated state	Providing presence (regime/therapy), Interventional procedure	
INDICATION: , Iron deficiency anemia.,PR	Skin appearance normal (finding), P	Anemia, Iron deficiency anemia, Deficien	Primary operation, Sedation procedure,	Intravenous sedation (procedure),
PREOPERATIVE DIAGNOSIS: , Right inguin	Pre-op diagnosis, Infertility, Post-op	Posterior pituitary disease, Retinitis Pigme	Repair of inguinal hernia, Tissue Dissecti	Sponge Dosage Form
PREOPERATIVE DIAGNOSES: , Colon cance	Skin appearance normal (finding), P	Colonic Polyps, Hemorrhoids, Diverticulos	Cauterization - action, Screening for cancer, Screening colonoscopy, Interve	

Step 3: Obtain the Final Dataset: Pre-Processing (Python)

The dataset that we created with UMLS concepts has transcription, diseases, symptoms, medications and procedures.

	transcription	symptoms	diseases	procedures	medications
0	SUBJECTIVE:, This 23-year-old white female pr...	Blood Pressure, Lymphadenopathy, Pressure (fin...	Rhinitis, Infantile Neuroaxonal Dystrophy, All...	Weighing patient, Transcranial magnetic stimul...	NaN
1	DESCRIPTION:,1. Normal cardiac chambers size....	effusion, Ejection as a Sports activity, Skin ...	Tricuspid Valve Insufficiency, Pericardial eff...	Doppler studies	NaN
2	PREOPERATIVE DIAGNOSIS: , Right inguinal herni...	Pre-op diagnosis, Pain, Infertility, Dressing-...	Retinitis Pigmentosa, Hernia sac, Hernia, Ingu...	Spinal Anesthesia, Dressing of skin or wound, ...	Solution Dosage Form, Sponge Dosage Form
3	PREOPERATIVE DIAGNOSIS: , Possible inflammator...	Able (finding), Pre-op diagnosis, Post-op diag...	Disease, Inflammatory Bowel Diseases, Intestin...	Sedation procedure, Interventional procedure, ...	NaN

Mtsamples dataset had the same transcription from where we initially took the transcriptions for our data preparation.

Unnamed: 0	description	medical_specialty	sample_name	transcription	keywords
0	A 23-year-old white female presents with comp...	Allergy / Immunology	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white female pr...	allergy / immunology, allergic rhinitis, aller...
1	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 2	PAST MEDICAL HISTORY:, He has difficulty climb...	bariatrics, laparoscopic gastric bypass, weigh...
2	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 1	HISTORY OF PRESENT ILLNESS: , I have seen ABC ...	bariatrics, laparoscopic gastric bypass, heart...
3	2-D M-Mode. Doppler.	Cardiovascular / Pulmonary	2-D Echocardiogram - 1	2-D M-MODE: , ,1. Left atrial enlargement wit...	cardiovascular / pulmonary, 2-d m-mode, dopple...

For training purposes, we merged our dataset with mtsamples dataset to get the medical specialty from mtsamples. After merging, there were many duplicate transcriptions which were removed. The datasets ‘mtsamples’ and ‘customMedicalDataset’ both had 4999 transcriptions, however many transcriptions were duplicate and hence we removed the repeated transcription. We were left with 2358 transcriptions. We dropped the ‘transcription’ column as it was no longer required. Hence our final dataset has the following columns: diseases, symptoms, procedures, medications and medical specialty.

Final Merged Dataset:

	symptoms	diseases	procedures	medications	medical_specialty
0	Blood Pressure, Lymphadenopathy, Pressure (fin...	Rhinitis, Infantile Neuroaxonal Dystrophy, All...	Weighing patient, Transcranial magnetic stimul...	NaN	Allergy / Immunology
1	Blood Pressure, Knee pain, Rectal hemorrhage, ...	Pseudotumor, Disease, Venous Insufficiency, Sl...	Weighing patient, Joint Examination, Knee, Rec...	Rectal Dosage Form	Bariatrics
2	Knee pain, Does play, Rectal hemorrhage, Ankle...	Pseudotumor, Disease, Venous Insufficiency, De...	Laparoscopic bypass of stomach, Gastric Bypass...	Rectal Dosage Form, Pack	Bariatrics
3	Pulmonary Valve Insufficiency, effusion, Press...	Tricuspid Valve Insufficiency, Pericardial eff...	Doppler studies, Enlargement procedure	NaN	Cardiovascular / Pulmonary

Further we checked for any missing or null values in symptoms, diseases and procedure. It showed that there were 73 null values for symptoms, 180 null values for diseases and 157

null values for procedure. We dropped all the null values which gave us 2107 unique transcriptions.

We had 39 different medical specialties in our dataset, however the count for almost 29 medical specialty was below 40. So, for better model efficiency, we decided to keep the top 10 medical specialties and perform classification for them. This left us with 1769 rows and the top 10 medical specialty from the dataset.

```
Medical Specialty having count more tha 40:
Medcial Specialty                Count
 Surgery                          927
 Radiology                        191
 Urology                          141
 General Medicine                 138
 SOAP / Chart / Progress Notes    127
 Neurology                         56
 Consult - History and Phy.        52
 Pediatrics - Neonatal             49
 Orthopedic                        45
 Psychiatry / Psychology           43
```

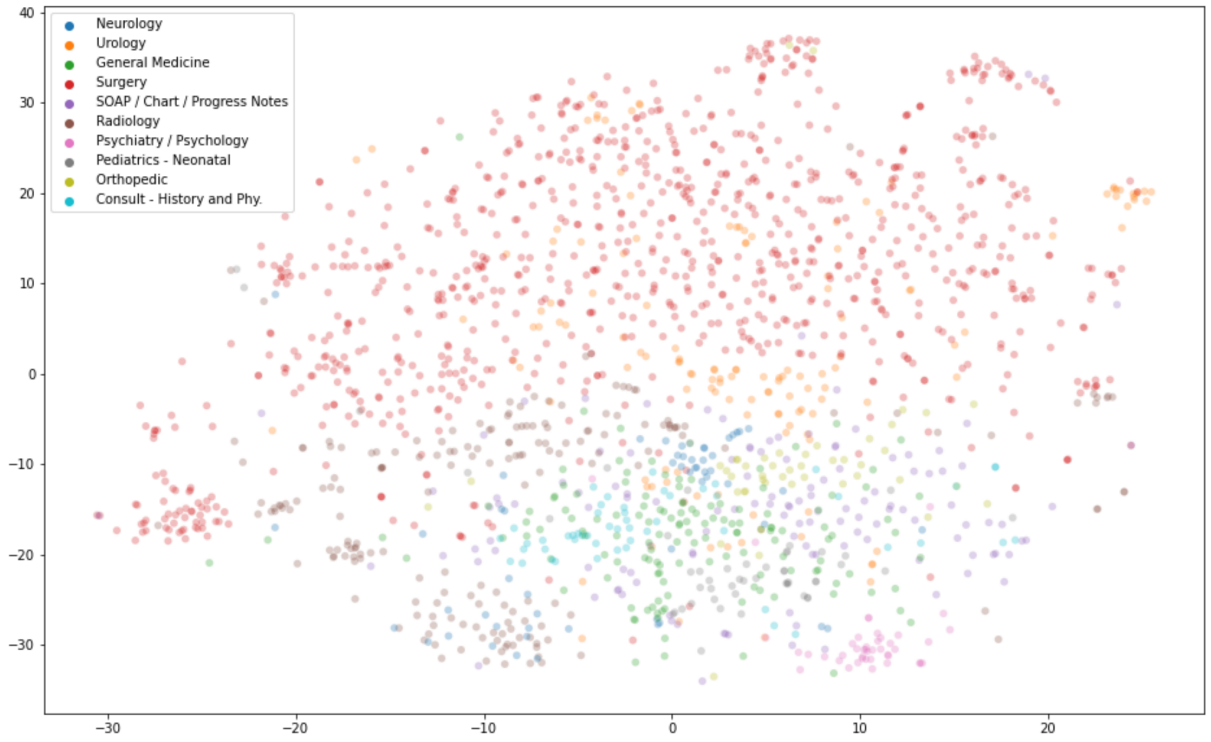
Using NLTK, we first cleaned the dataset by removing the special characters and punctuations. We later lemmatized the dataset to get the base or the root form for the words. Finally we combined our symptoms, diseases and procedures into one column called ‘alldata’. We omitted the ‘medications’ column for NLTK processing as it contained a lot of null values and was not quite useful.

alldata	medical_specialty
headache gait functional disorder medical h...	Neurology
sharp sensation quality dressing activity of ...	Urology
hypersensitivity illness finding medical h...	General Medicine

Modelling

For feature extraction, we used TfidVectorizer from sklearn to generate 2000 tf-idf features. It converts a collection of raw documents to a matrix of TF-IDF features. TFIDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic to show how important a word is to a document. The tf-idf value increases with respect to the number of times a word appears in the document.

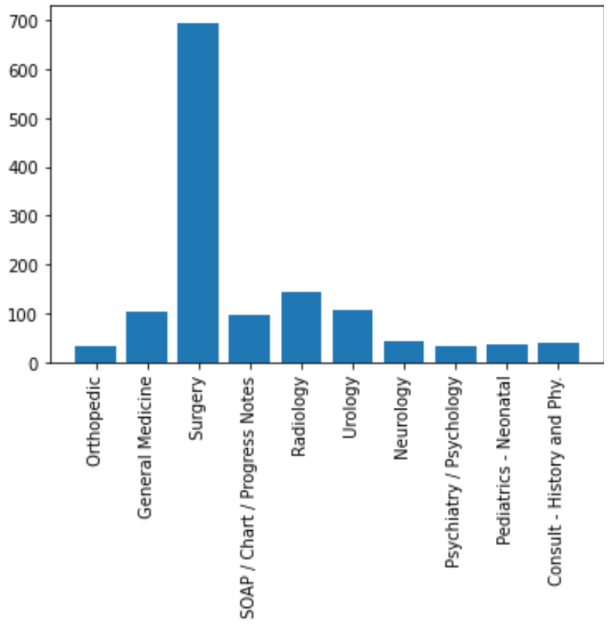
We then used TSNE to visualize the tf-idf features for the medical specialty. TSNE stands for T-distributed Stochastic Neighbor Embedding which is a machine learning algorithm for visualization. It works well for high-dimensional data in a two or three low-dimensional space for visualisation. The t-sne plot here shows us that there are few medical specialties that are overlapping.



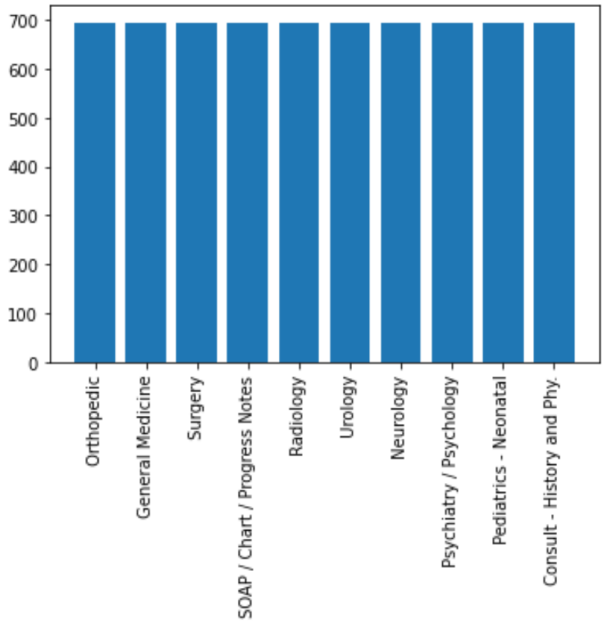
We performed Principal Component Analysis to reduce the dimensionality of the features. n_components was set to 0.99 after which we had reduced the feature count from 2000 to 958. We then did a train-test split using the stratify option, so that the split is performed with the same proportion of medical specialties.

Our Multiclass dataset was highly imbalanced as shown below. The below graph(left) shows that 52% of the data had the medical specialty as 'Surgery'. Therefore to balance the train dataset, we used SMOTE for oversampling the X_train and y_train.

Before SMOTE:



After SMOTE:



For modelling, we used several models such as SVM, Logistic Regression, Decision Tree KNN and Naive Bayes and checked the accuracy and f1 scores.

Results

Midterm Results: We were able to create our custom dataset from the raw transcriptions. We installed and learned the Apache cTAKES tool to process raw clinical data. Finally, on using cTAKES along with some JAVA programs we were able to generate the custom dataset containing some relevant features. The initial dataset had only 30 records since it was created for the purpose of testing and understanding different tools. Later on we expanded the dataset to 4999 records.

Final Presentation Results: We tried multiple models which included Logistic Regression, Support Vector Machine, Decision Tree classifier, KNN classifier and Naive Bayes for multiclass classification. We observed that SVM gave us the best results with accuracy of 0.80 and F1 score of 0.64 (highlighted in orange).

Sr No.	PCA	SMOTE	MODEL	Fine Tuning Parameters	Accuracy	F1 Score
1	✔	✖	Decision Tree	✖	0.57	0.20
2	✔	✖	KNN	✖	0.66	0.51
3	✔	✖	Logistic Regression	✖	0.77	0.54
4	✔	✖	SVM	✖	0.80	0.64

Final Report Results: Further, we balanced the training dataset using SMOTE and also adjusted some configuration parameters such as random_state, PCA n_components, tfidf Vectorizer max_ features. Also, we removed some records containing null values for symptoms, diseases and procedures. After performing the above steps, we observed a significant improvement in the results for some models. This time, instead of SVM, the Logistics Regression model gave the best accuracy (0.86) and f1 score (0.75) as highlighted in the table below.

Attempt No.	PCA	SMOTE	MODEL	Fine Tuning Parameters	Accuracy	F1 Score
1	✔	✖	Decision Tree	✖	0.57	0.20
2	✔	✖	KNN	✖	0.66	0.51
3	✔	✖	Logistic Regression	✖	0.77	0.54

4	✔	✖	SVM	✖	0.80	0.64
5	✔	✔	Naive Bayes	✖	0.52	0.16
6	✔	✔	SVM	✔	0.83	0.69
7	✔	✔	Logistic Regression	✔	**0.86**	**0.75**

Scores for each medical specialty:

	precision	recall	f1-score	support
Neurology	0.64	0.50	0.56	14
Urology	0.79	0.86	0.82	35
General Medicine	0.69	0.63	0.66	35
Surgery	0.97	0.94	0.96	232
SOAP / Chart / Progress Notes	0.71	0.69	0.70	32
Radiology	0.77	0.90	0.83	48
Psychiatry / Psychology	0.91	0.91	0.91	11
Pediatrics - Neonatal	0.70	0.58	0.64	12
Orthopedic	0.73	0.73	0.73	11
Consult - History and Phy.	0.61	0.85	0.71	13
accuracy			0.86	443
macro avg	0.75	0.76	0.75	443
weighted avg	0.86	0.86	0.86	443

Conclusion

We researched on how Knowledge based WSD(using UMLS) can assist in analyzing raw clinical texts and extract important medical concepts such as symptoms, diseases and procedures. This information is then further used to perform multi-class classification such as predicting the medical specialty based on the transcription/clinical text. We also saw that cTAKES is an efficient way to extract useful biomedical concepts from the raw clinical data.

Future Work

This research could be helpful in creating a Web/Mobile Application that can predict the medical speciality based on the given transcription. Such applications could be used in hospitals to refer the patient to the correct department. Also, based on a similar model, an AI based chatbot can be developed that performs classification on symptoms and recommends the medical specialty.

References

Elhadad, N. (2013, December). *Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts*. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. https://www.sciencedirect.com/science/article/pii/S1532046413001196?via%3Dihub

Garla, V., & Brandt, C. (2012, October 16). *Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification*. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. https://academic.oup.com/jamia/article/20/5/882/728540

Godinez, E., Szlávik, Z., Contempré, E., & Sips, R.-J. (n.d.). *What do You Mean, Doctor? A Knowledge-based Approach for Word Sense Disambiguation of Medical Terminology*. What do You Mean, Doctor? A Knowledge-based Approach for Word Sense Disambiguation of Medical Terminology. https://www.scitepress.org/Link.aspx?doi=10.5220/0010180502730280

Jimeno-Yepes, A. J., & Aronson, A. R. (2010, November 22). *Knowledge-based biomedical word sense disambiguation: comparison of approaches*. Knowledge-based biomedical word sense disambiguation: comparison of approaches. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-569

Long, A. (2018). *Introduction to Clinical Natural Language Processing : Predicting Hospital Readmission with Discharge Summaries*. Introduction to Clinical Natural Language Processing. https://towardsdatascience.com/introduction-to-clinical-natural-language-processing-predicting-hospital-readmission-with-1736d52bc709

MTSamples. (n.d.). MTSamples. https://www.mtsamples.com/

Pesaranghader, A., Matwin, S., Sokolova, M., & Pesaranghader, A. (2019, February 26). *deepBioWSD: effective deep neural word sense disambiguation of biomedical text data*. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data.

https://doi.org/10.1093/jamia/ocy189

Tancev, G. (2019, October 25). *Mining and Classifying Medical Documents*. Mining and Classifying Medical Documents. https://towardsdatascience.com/mining-and-classifying-medical-text-documents-1876462f73bc

UMLS. (n.d.). *Apache cTAKES*. Apache cTAKES. https://github.com/mccullen/ctakes-example/wiki/User-Tutorial

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019, January 07). *A clinical text classification paradigm using weak supervision and deep representation*. A clinical text classification paradigm using weak supervision and deep representation. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0723-6

Wang, Y., Wang, M., & Fujita, H. (2020, February 29). *Word Sense Disambiguation: A comprehensive knowledge exploitation framework*. Word Sense Disambiguation: A comprehensive knowledge exploitation framework. https://www.sciencedirect.com/science/article/pii/S0950705119304344?via%3Dihub

Weng, W.-H., Wagholikar, K. B., McCray, A., Szolovits, P., & Chueh, H. C. (2017, December 01). *Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach*. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0556-8

Yepes, A. J. (2017, September). *Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation*. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. https://www.sciencedirect.com/science/article/pii/S1532046417301806?via%3Dihub

References

Elhadad, N. (2013, December). *Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts*. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. https://www.sciencedirect.com/science/article/pii/S1532046413001196?via%3Dihub

Garla, V., & Brandt, C. (2012, October 16). *Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification*. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. https://academic.oup.com/jamia/article/20/5/882/728540

Godinez, E., Szlávik, Z., Contempré, E., & Sips, R.-J. (n.d.). *What do You Mean, Doctor? A Knowledge-based Approach for Word Sense Disambiguation of Medical Terminology*. What do You Mean, Doctor? A Knowledge-based Approach for Word Sense Disambiguation of Medical Terminology. https://www.scitepress.org/Link.aspx?doi=10.5220/0010180502730280

Jimeno-Yepes, A. J., & Aronson, A. R. (2010, November 22). *Knowledge-based biomedical word sense disambiguation: comparison of approaches*. Knowledge-based biomedical word sense disambiguation: comparison of approaches. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-569

Long, A. (2018). *Introduction to Clinical Natural Language Processing : Predicting Hospital Readmission with Discharge Summaries*. Introduction to Clinical Natural Language

Processing. https://towardsdatascience.com/introduction-to-clinical-natural-language-processing-predicting-hospital-readmission-with-1736d52bc709

MTSamples. (n.d.). MTSamples. https://www.mtsamples.com/

Pesaranghader, A., Matwin, S., Sokolova, M., & Pesaranghader, A. (2019, February 26). *deepBioWSD: effective deep neural word sense disambiguation of biomedical text data*. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. https://doi.org/10.1093/jamia/ocy189

Tancev, G. (2019, October 25). *Mining and Classifying Medical Documents*. Mining and Classifying Medical Documents. https://towardsdatascience.com/mining-and-classifying-medical-text-documents-1876462f73bc

UMLS. (n.d.). *Apache cTAKES*. Apache cTAKES. https://github.com/mccullen/ctakes-example/wiki/User-Tutorial

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019, January 07). *A clinical text classification paradigm using weak supervision and deep representation*. A clinical text classification paradigm using weak supervision and deep representation. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0723-6

Wang, Y., Wang, M., & Fujita, H. (2020, February 29). *Word Sense Disambiguation: A comprehensive knowledge exploitation framework*. Word Sense Disambiguation: A comprehensive knowledge exploitation framework.

References

Elhadad, N. (2013, December). *Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts*. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. https://www.sciencedirect.com/science/article/pii/S1532046413001196?via%3Dihub

Garla, V., & Brandt, C. (2012, October 16). *Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification*. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. https://academic.oup.com/jamia/article/20/5/882/728540

Godinez, E., Szlávik, Z., Contempré, E., & Sips, R.-J. (n.d.). *What do You Mean, Doctor? A Knowledge-based Approach for Word Sense Disambiguation of Medical Terminology*. What do You Mean, Doctor? A Knowledge-based Approach for Word Sense Disambiguation of Medical Terminology. https://www.scitepress.org/Link.aspx?doi=10.5220/0010180502730280

Jimeno-Yepes, A. J., & Aronson, A. R. (2010, November 22). *Knowledge-based biomedical word sense disambiguation: comparison of approaches*. Knowledge-based biomedical word sense disambiguation: comparison of approaches. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-569

Long, A. (2018). *Introduction to Clinical Natural Language Processing : Predicting Hospital Readmission with Discharge Summaries*. Introduction to Clinical Natural Language Processing. https://towardsdatascience.com/introduction-to-clinical-natural-language-processing-predicting-hospital-readmission-with-1736d52bc709

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Collados, J. C. (2021, July 13). *Analysis and Evaluation of Language Models for Word Sense Disambiguation*. Analysis and Evaluation of Language Models for Word Sense Disambiguation. https://direct.mit.edu/coli/article/47/2/387/98520/Analysis-and-Evaluation-of-Language-Models-for

MTSamples. (n.d.). MTSamples. https://www.mtsamples.com/

Pesaranghader, A., Matwin, S., Sokolova, M., & Pesaranghader, A. (2019, February 26). *deepBioWSD: effective deep neural word sense disambiguation of biomedical text data*. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. https://doi.org/10.1093/jamia/ocy189

Sabbir, A., Yepes, A. J., & Kavuluru, R. (2017, October). *Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings*. Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5792196/

Tancev, G. (2019, October 25). *Mining and Classifying Medical Documents*. Mining and Classifying Medical Documents. https://towardsdatascience.com/mining-and-classifying-medical-text-documents-1876462f73bc

UMLS. (n.d.). *Apache cTAKES*. Apache cTAKES. https://github.com/mccullen/ctakes-example/wiki/User-Tutorial

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019, January 07). *A clinical text classification paradigm using weak supervision and deep representation*. A clinical text classification paradigm using weak supervision and deep representation.

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0723-6

Wang, Y., Wang, M., & Fujita, H. (2020, February 29). *Word Sense Disambiguation: A comprehensive knowledge exploitation framework*. Word Sense Disambiguation: A comprehensive knowledge exploitation framework. https://www.sciencedirect.com/science/article/pii/S0950705119304344?via%3Dihub

Weng, W.-H., Wagholikar, K. B., McCray, A., Szolovits, P., & Chueh, H. C. (2017, December 01). *Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach*. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0556-8

Yepes, A. J. (2017, September). *Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation*. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. https://www.sciencedirect.com/science/article/pii/S1532046417301806?via%3Dihub

https://www.sciencedirect.com/science/article/pii/S0950705119304344?via%3Dihub

Weng, W.-H., Wagholikar, K. B., McCray, A., Szolovits, P., & Chueh, H. C. (2017, December 01). *Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach*. Medical subdomain classification of clinical

notes using a machine learning-based natural language processing approach. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0556-8

Yepes, A. J. (2017, September). *Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation*. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. https://www.sciencedirect.com/science/article/pii/S1532046417301806?via%3Dihub