TEXT CLASSIFICATION OF MEDICAL TRANSCRIPTS

December 12th ,2022

Kavya Mistry, MS Data Science

1. INTRODUCTION

In modern era, almost every domain has large amount of data. One of the most common is text data. The quantity of complicated documents and texts that demand a greater grasp of machine learning technologies to effectively identify texts in numerous applications has recently increased exponentially. Many machine learning algorithms have outperformed in Natural Language Processing (NLP). The ability of these learning algorithms to recognize complicated models and non-linear correlations within data is critical to their effectiveness. [1] Many NLP experts have leveraged Text Classification methods for applications such as email classification, sentiment analysis, particularly in healthcare sector it's booming. Classification of clinical documentation, transcribing patient interactions and conducting conversational AI, analysing unstructured clinical notes, practice in Electronic Health Records (EHR) and so on are instances of what natural language processing can achieve. [2]

The majority of text classification and document categorization systems may be broken down into four stages: feature extraction, dimension reductions, classifier selection, and assessments.[1] The aim of this project is to classify the medical transcriptions dataset into the speciality labels such as surgery, cardiovascular, neurology, etc.

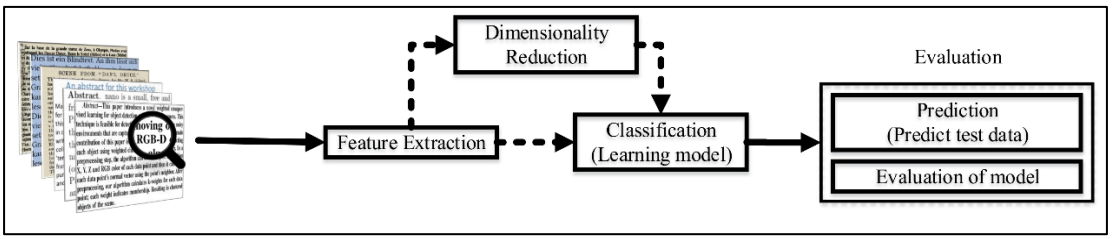

Figure 1- Flowchart for Text classification. Image Source: Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text Classification Algorithms: A Survey. Information. 2019; 10(4):150. https://doi.org/10.3390/info10040150

2. PROBLEM DEFINITION

The task is to train and test the model on an offline dataset from Kaggle to classify target variable or dependent variable which is medical speciality label. The independent variable which will be focused on is transcription column which has text relating to the patient's health.

The important questions or problem statements are:

i. How should I prepare/clean the dataset at hand, so our prediction rate is efficient?
ii. What model should I choose and on what basis?

iii. What configuration parameters should we use for our model?
iv. What should I use for the feature engineering task?
v. If any visualization plots will be useful for analyses?
vi. How to evaluate the trained models?

3. BACKGROUND

The classification of clinical text is a crucial challenge in medical natural language processing. Existing research has traditionally relied on rules or knowledge sources-based feature engineering, but only a few studies have taken use of deep learning techniques' excellent representation learning capabilities. Clinical records are an essential sort of electronic health record (EHR) data because they frequently contain extensive and significant patient information as well as doctors' clinical experiences. Text categorization, as a fundamental job of natural language processing, is crucial in healthcare record retrieval and organization; it may also aid clinical decision making and cohort identification.[3]

Google developers have access to a wide range of data preparation and model design choices as a result of decades of study. The availability of a very vast variety of plausible choices to pick from, but on the other hand, considerably increases the complexity and breadth of the specific situation at hand. Given that the optimal solutions may not be evident, a naïve strategy would be to exhaust all available options, cutting some by intuition. However, that would be too costly.[4]

They wanted to make the process of picking a text categorization model as simple as possible. Their aim for a given dataset is to discover the technique that delivers close to maximum accuracy while using the least amount of computing time for training. They did 450K tests across challenges of various sorts (particularly sentiment analysis and topic classification tasks), utilizing 12 datasets and rotating between different data pre-treatment approaches and model architectures for each dataset. This assisted them in identifying dataset factors that impact optimum decisions.[4]

The experimentation is summarized in the model selection process and flowchart below.

Algorithm for Data Preparation and Model Building

1. Calculate the number of samples/number of words per sample ratio.
2. If this ratio is less than 1500, tokenize the text as n-grams and use a simple multi-layer perceptron (MLP) model to classify them (left branch in the flowchart below):
   a. Split the samples into word n-grams; convert the n-grams into vectors.
   b. Score the importance of the vectors and then select the top 20K using the scores.
   c. Build an MLP model.
3. If the ratio is greater than 1500, tokenize the text as sequences and use a sepCNN model to classify them (right branch in the flowchart below):
   a. Split the samples into words; select the top 20K words based on their frequency.
   b. Convert the samples into word sequence vectors.
   c. If the original number of samples/number of words per sample ratio is less than 15K, using a fine-tuned pre-trained embedding with the sepCNN model will likely provide the best results.
4. Measure the model performance with different hyperparameter values to find the best model configuration for the dataset.

Figure 2-Algorithm for Text Classification. Image Source: Step 2.5: Choose a Model | Machine Learning | Google Developers. Google Developers. Accessed December 11, 2022. https://developers.google.com/machine-learning/guides/text-classification/step-2-5

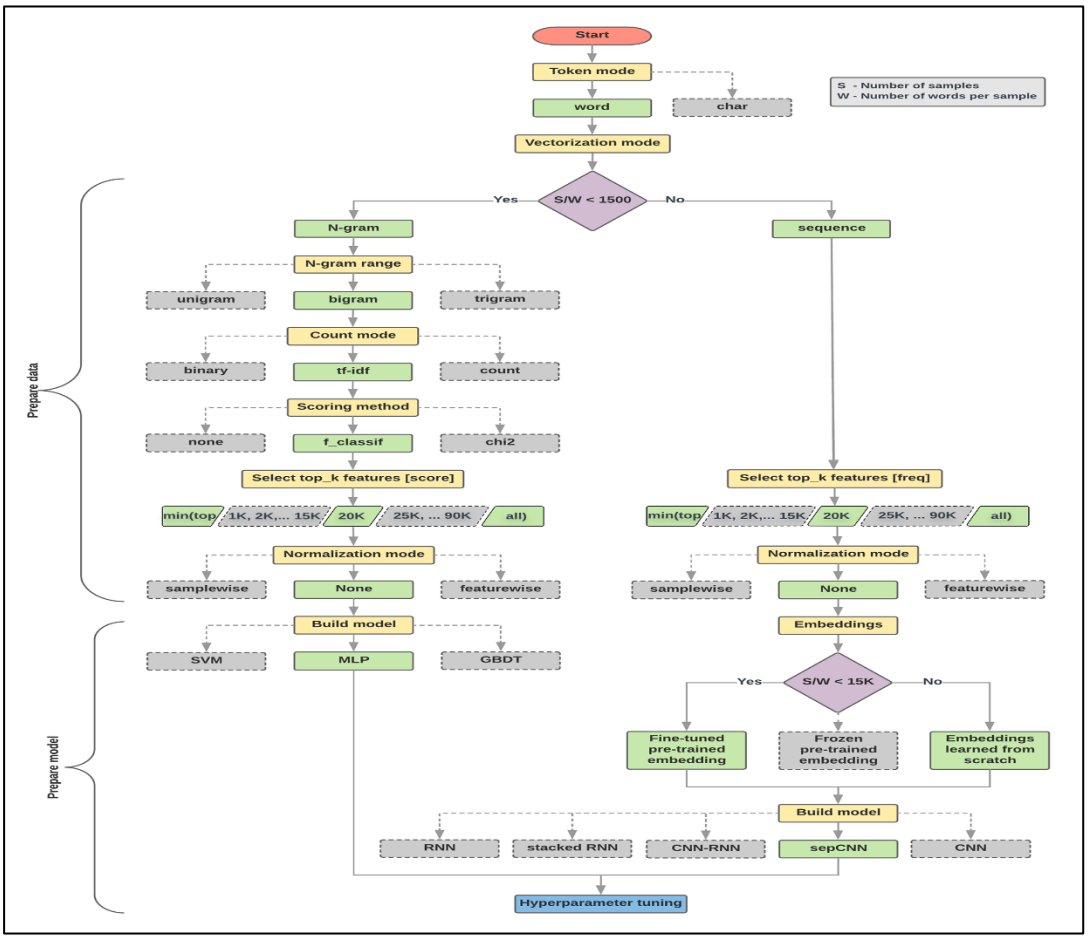

Figure 3-Flowchart Of the Algorithm Image Source: Step 2.5: Choose a Model | Machine Learning | Google Developers. Google Developers. Accessed December 11,

2022. https://developers.google.com/machine-learning/guides/text-classification/step-2-5

4. ABOUT THE DATASET

HIPAA privacy laws make it very difficult to find real-life medical data or having a free access. This dataset provides a remedy by supplying examples of medical transcripts. Sample medical transcriptions for several medical disciplines are included in this collection, I downloaded from Kaggle (csv format). The usability of the dataset is 8.53. There are total 6 columns and 4998 rows.[5] The attributes are as follows:

- ID (0,1, 2…)
- Description: Short description of transcription
- Medical specialty: Medical specialty classification of transcription(target)
- Transcription title
- Sample medical transcriptions: Detailed description about health concerns of patient
- Relevant keywords from transcription

All variables are string except for ID column which is an integer.

	description	medical_specialty	sample_name	transcription	keywords
0	A 23-year-old white female presents	Allergy / Immunology	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white fer	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra
1	Consult for laparoscopic gastric bypc	Bariatrics	Laparoscopic Gastric Bypass Consult -	PAST MEDICAL HISTORY:, He has difficul	bariatrics, laparoscopic gastric bypass, weight loss programs, gastric bypass, atkin's diet, weight watcher's
2	Consult for laparoscopic gastric bypc	Bariatrics	Laparoscopic Gastric Bypass Consult -	HISTORY OF PRESENT ILLNESS:, I have s	bariatrics, laparoscopic gastric bypass, heart attacks, body weight, pulmonary embolism, potential compli
3	2-D M-Mode. Doppler.	Cardiovascular / Pulm	2-D Echocardiogram - 1	2-D M-MODE: , ,1. Left atrial enlargeme	cardiovascular / pulmonary, 2-d m-mode, doppler, aortic valve, atrial enlargement, diastolic function, eje
4	2-D Echocardiogram	Cardiovascular / Pulm	2-D Echocardiogram - 2	1. The left ventricular cavity size and wa	cardiovascular / pulmonary, 2-d, doppler, echocardiogram, annular, aortic root, aortic valve, atrial, atrium
5	Morbid obesity. Laparoscopic antec	Bariatrics	Laparoscopic Gastric Bypass	PREOPERATIVE DIAGNOSIS:, Morbid obe	bariatrics, gastric bypass, eea anastomosis, roux-en-y, antegastric, antecolic, morbid obesity, roux limb, g
6	Liposuction of the supraumbilical	Bariatrics	Liposuction	PREOPERATIVE DIAGNOSES:,1. Deformi	bariatrics, breast reconstruction, excess, lma anesthesia, lipodystrophy, liposuction, abdomen, drain site,
7	2-D Echocardiogram	Cardiovascular / Pulm	2-D Echocardiogram - 3	2-D ECHOCARDIOGRAM,Multiple views c	cardiovascular / pulmonary, 2-d echocardiogram, cardiac function, doppler, echocardiogram, multiple vie
8	Suction-assisted lipectomy - lipodyst	Bariatrics	Lipectomy - Abdomen/Thighs	PREOPERATIVE DIAGNOSIS:, Lipodystro	bariatrics, lipodystrophy, abd pads, suction-assisted lipectomy, abdomen, aspirate, lipectomy, perineum, s
9	Echocardiogram and Doppler	Cardiovascular / Pulm	2-D Echocardiogram - 4	DESCRIPTION:,1. Normal cardiac chamb	cardiovascular / pulmonary, ejection fraction, lv systolic function, cardiac chambers, regurgitation, tricusp
10	Morbid obesity. Laparoscopic Roux-	Bariatrics	Laparoscopic Gastric Bypass - 1	PREOPERATIVE DIAGNOSIS:, Morbid obe	bariatrics, morbid obesity, roux-en-y, gastric bypass, antecolic, antegastric, anastamosis, esophagogastro
11	Normal left ventricle, moderate biat	Cardiovascular / Pulm	2-D Doppler	2-D STUDY,1. Mild aortic stenosis, wideh	cardiovascular / pulmonary, 2-d study, doppler, tricuspid regurgitation, heart pressures, stenosis, ventricul
12	Cerebral Angiogram - moyamoya dis	Neurology	Moyamoya Disease	CC:, Confusion and slurred speech.,HX , (primarily obtained from boyfriend): This 31 y/o RHF experienced a "flu-like illness 6-8 weeks prior to prese
13	Patient presented to the bariatric su	Bariatrics	Gastric Bypass Discussion - 3	PAST MEDICAL HISTORY:, Significant for	bariatrics, weight watchers, roux en y, atkins, medifast, meridia, south beach, cabbage, diets, laparoscopic
14	Surgical removal of completely bon	Dentistry	Bony Impacted Teeth Removal	PREOPERATIVE DIAGNOSIS:, Completely	dentistry, intraoral, bony impacted teeth, throat pack, buccal aspect, saline solution, gut sutures, envelop
15	Preoperative visit for weight manage	Bariatrics	Laparoscopic Gastric Banding - Preop	HISTORY OF PRESENT ILLNESS:, I have se	bariatrics, laparoscopic gastric banding, pulmonary embolism, lap banding, potential complications, gastri
16	Neck exploration; tracheostomy; ur	Cardiovascular / Pulm	Tracheostomy	PREOPERATIVE DIAGNOSES,Airway obst	cardiovascular / pulmonary, airway, laryngology, shiley, alteration of voice, bronchi, bronchoscopy, cannu

Figure 4- Dataset

There are some null values in the dataset as well.

5. NLP APPROACH

The action plan for the project, is as follows:

1. **Studying the dataset**: I analysed the data, to check which columns are relevant for predicting the target variable. From the 6 columns, Transcription Description is the most useful column to predict the medical speciality. I studied the recent techniques used for health data classification, NLP algorithms used and challenges in the problem statement.
2. **Data Cleaning and EDA**: The libraries used for this project is pandas, numpy, NLTK, sklearn, matplotlib, Regular Expression, OS, and so on. For pre-processing the text data, I implemented a function for removing duplicate instances, replace spaces with underscore (if any), lower case the titles for columns. Also, I displayed word count of the medical speciality and transcriptions. I computed sample size of the medical speciality. I dealt with null values by removing them.
3. **Text Normalisation:** This step is performed on the transcription column. The tasks done were converting the transcription text to lower Case, removing punctuation and numbers, tokenisation of the transcription, Lemmatisation, remove stop words and dropping redundant columns.
4. **Feature Extraction using N-grams:** For this, sklearn class, Count Vectorizer was used to develop matrix of n-grams features from transcriptions. And as N-grams can be applied on list not on data frame, I used flatten function to do the necessary. Unigram. Unigram-bigram, bigram, bigram-trigram, trigram features were extracted. Additionally, dimensions of each feature vector were generated.
5. **Building the classification Model:** The data was split in training and testing data. Three supervised learning classifiers were trained on each of the n-gram feature representations, namely K-Nearest Neighbours, Decision Tree, and Random Forest. A function was created to evaluate the performance using F-1, Recall and Precision. I also visualized classification prediction to display and analyse best method. Moreover, I performed Dimensionality reduction for better outputs.
6. **Testing and Evaluation:** I obtained best classifier and feature vector, tested it, evaluated the class labels and predicted the results.

6. IMPLEMENTATION

For Text Pre-processing:

Word Counts of Each Medical Specialty

In [5]:
```
medical_specialty_list = [] ; word_count_list =[]
for medical_specialty in df['medical_specialty'].unique():
    df_filter = df.loc[(df['medical_specialty'] == medical_specialty)]
    word_count_temp = df_filter['transcription'].str.split().str.len().sum()
    medical_specialty_list.append(medical_specialty)
    word_count_list.append(word_count_temp)
word_count_df = pd.DataFrame({'Medical Specialty':medical_specialty_list, 'Word Count':word_count_list})
word_count_df['Word Count'] = word_count_df['Word Count'].astype('int')
word_count_df = word_count_df.sort_values('Word Count', ascending=False)
word_count_df.reset_index(drop=True)
```

Out[5]:

	Medical Specialty	Word Count
0	Neurosurgery	54233
1	Discharge Summary	43103
2	ENT - Otolaryngology	42032

In [6]:
```
total_word_count = df['transcription'].str.split().str.len().sum()
print(f'The word count of all transcription is: {int(total_word_count)}')
```

```
The word count of all transcription is: 139368
```

Figure 5- Word count of Medical Speciality

Sample Size of Each Medical Specialty

In [7]:
```
vc(df, 'medical_specialty')
```

```
STATUS: Value counts of "medical_specialty"...
```

medical_specialty	count	percentage (%)
Discharge Summary	108	36.000000
ENT - Otolaryngology	98	32.666667
Neurosurgery	94	31.333333

Figure 6- Sample size of medical speciality

In [9]:
```
# to print data shape
print(f'data shape is: {df.shape}')

# to identify the null values by descending order
df.isnull().sum().sort_values(ascending = False)
```

```
data shape is: (300, 5)
```

Out[9]:
```
keywords             56
transcription         2
sample_name           0
medical_specialty     0
description           0
dtype: int64
```

Figure 7- Detecting null values

For Text Normalisation:

In [19]:
```
# To convert transcription into lowercase
def lower(df, attribute):
    df.loc[:,attribute] = df[attribute].apply(lambda x : str.lower(x))
    return df
df = lower(df,'transcription')
df.head(3)
```

Out[19]:

	medical_specialty	transcription
2656	Neurosurgery	title of operation:, a complex closure and debridement of wound.,indication for surgery:, the patient is a 26-year-old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant, just below the...
2657	Neurosurgery	title of operation: , placement of right new ventriculoperitoneal (vp) shunts strata valve and to removal of right frontal ommaya reservoir.,indication for surgery: , the patient is a 2-month-old infant, born premature with intraventricular hemorrhage...
2658	Neurosurgery	preoperative diagnosis: , aqueductal stenosis .postoperative diagnosis:, aqueductal stenosis .title of procedure: ,endoscopic third ventriculostomy.,anesthesia: , general endotracheal tube anesthesia.,devices:, bactiseal ventricular catheter with a...

Figure 8- Convert text of transcriptions to lowercase

In [20]:
```
# To remove transcription punctuation and numbers

warnings.filterwarnings('ignore')
def remove_punc_num(df, attribute):
    df.loc[:,attribute] = df[attribute].apply(lambda x : " ".join(re.findall('[\w]+',x)))
    df[attribute] = df[attribute].str.replace('\d+', '')
    return df
df =remove_punc_num(df, 'transcription')
df_no_punc =df.copy()
df.head(3)
```

Out[20]:

	medical_specialty	transcription
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...
2658	Neurosurgery	preoperative diagnosis aqueductal stenosis postoperative diagnosis aqueductal stenosis title of procedure endoscopic third ventriculostomy anesthesia general endotracheal tube anesthesia devices bactiseal ventricular catheter with an aesculap burr hol...

Figure 9- Remove punctuations

In [21]:
```
# to tokenise transcription

# import nltk
tk =WhitespaceTokenizer()
def tokenise(df, attribute):
    df['tokenised'] = df.apply(lambda row: tk.tokenize(str(row[attribute])), axis=1)
    return df
df =tokenise(df, 'transcription')
df_experiment =df.copy()
df.head(3)
```

Out[21]:

	medical_specialty	transcription	tokenised
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...	[title, of, operation, a, complex, closure, and, debridement, of, wound, indication, for, surgery, the, patient, is, a, year, old, female, with, a, long, history, of, shunt, and, hydrocephalus, presenting, with, a, draining, wound, in, the, right, upp...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...	[title, of, operation, placement, of, right, new, ventriculoperitoneal, vp, shunts, strata, valve, and, to, removal, of, right, frontal, ommaya, reservoir, indication, for, surgery, the, patient, is, a, month, old, infant, born, premature, with, intra...
2658	Neurosurgery	preoperative diagnosis aqueductal stenosis postoperative diagnosis aqueductal stenosis title of procedure endoscopic third ventriculostomy anesthesia general endotracheal tube anesthesia devices bactiseal ventricular catheter with an aesculap burr hol...	[preoperative, diagnosis, aqueductal, stenosis, postoperative, diagnosis, aqueductal, stenosis, title, of, procedure, endoscopic, third, ventriculostomy, anesthesia, general, endotracheal, tube, anesthesia, devices, bactiseal, ventricular, catheter, w...

Figure 10- Tokenisation

In [22]:
```
from nltk.stem.snowball import SnowballStemmer
def stemming(df, attribute):
    # Use English stemmer.
    stemmer = SnowballStemmer("english")
    df['stemmed'] = df[attribute].apply(lambda x: [stemmer.stem(y) for y in x]) # Stem every word.
    return df
df =stemming(df_experiment, 'tokenised')
df.head(2)
```

Out[22]:

	medical_specialty	transcription	tokenised	stemmed
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...	[title, of, operation, a, complex, closure, and, debridement, of, wound, indication, for, surgery, the, patient, is, a, year, old, female, with, a, long, history, of, shunt, and, hydrocephalus, presenting, with, a, draining, wound, in, the, right, upp...	[titl, of, oper, a, complex, closur, and, debrid, of, wound, indic, for, surgeri, the, patient, is, a, year, old, femal, with, a, long, histori, of, shunt, and, hydrocephalus, present, with, a, drain, wound, in, the, right, upper, quadrant, just, belo...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...	[title, of, operation, placement, of, right, new, ventriculoperitoneal, vp, shunts, strata, valve, and, to, removal, of, right, frontal, ommaya, reservoir, indication, for, surgery, the, patient, is, a, month, old, infant, born, premature, with, intra...	[titl, of, oper, placement, of, right, new, ventriculoperiton, vp, shunt, strata, valv, and, to, remov, of, right, frontal, ommaya, reservoir, indic, for, surgeri, the, patient, is, a, month, old, infant, born, prematur, with, intraventricular, hemorr...

Figure 11- Stemming

In [25]:
```
# Removing stop words
def remove_stop_words(df, attribute):
    stop = stopwords.words('english')
    df['stemmed_without_stop'] = df[attribute].apply(lambda x: ' '.join([word for word in x if word not in (stop)]))
    return df
df = remove_stop_words(df, 'stemmed')
df.head(2)
```

Out[25]:

	medical_specialty	transcription	tokenised	stemmed	stemmed_without_stop
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...	[title, of, operation, a, complex, closure, and, debridement, of, wound, indication, for, surgery, the, patient, is, a, year, old, female, with, a, long, history, of, shunt, and, hydrocephalus, presenting, with, a, draining, wound, in, the, right, upp...	[titl, of, oper, a, complex, closur, and, debrid, of, wound, indic, for, surgeri, the, patient, is, a, year, old, femal, with, a, long, histori, of, shunt, and, hydrocephalus, present, with, a, drain, wound, in, the, right, upper, quadrant, just, belo...	titl oper complex closur debrid wound indic surgeri patient year old femal long histori shunt hydrocephalus present drain wound right upper quadrant costal margin lanc general surgeri resolv howev continu drain evid fever crp normal shunt ct normal th...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...	[title, of, operation, placement, of, right, new, ventriculoperitoneal, vp, shunts, strata, valve, and, to, removal, of, right, frontal, ommaya, reservoir, indication, for, surgery, the, patient, is, a, month, old, infant, born, premature, with, intra...	[titl, of, oper, placement, of, right, new, ventriculoperiton, vp, shunt, strata, valv, and, to, remov, of, right, frontal, ommaya, reservoir, indic, for, surgeri, the, patient, is, a, month, old, infant, born, prematur, with, intraventricular, hemorr...	titl oper placement right new ventriculoperiton vp shunt strata valv remov right frontal ommaya reservoir indic surgeri patient month old infant born prematur intraventricular hemorrhag ommaya reservoir recommend remov replac new vp shunt preop diagno...

Figure 12- Removing stopwords

	medical_specialty	stemmed_without_stop	encoded_target
2656	Neurosurgery	titl oper complex closur debrid wound indic surgeri patient year old femal long histori shunt hydrocephalus present drain wound right upper quadrant costal margin lanc general surgeri resolv howev continu drain evid fever crp normal shunt ct normal th...	2
2657	Neurosurgery	titl oper placement right new ventriculoperiton vp shunt strata valv remov right frontal ommaya reservoir indic surgeri patient month old infant born prematur intraventricular hemorrhag ommaya reservoir recommend remov replac new vp shunt preop diagno...	2
2658	Neurosurgery	preoper diagnosi aqueduct stenosi postop diagnosi aqueduct stenosi titl procedur endoscop third ventriculostomi anesthesia general endotrach tube anesthesia devic bactis ventricular cathet aesculap burr hole port skin prepar chloraprep complic none sp...	2
2661	Neurosurgery	procedur placement left ventriculostomi via twist drill preoper diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur postop diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur indic proced...	2
2662	Neurosurgery	preoper diagnos increas intracrani pressur cerebr edema due sever brain injuri postop diagnos increas intracrani pressur cerebr edema due sever brain injuri procedur burr hole insert extern ventricular drain cathet anesthesia bedsid sedat procedur sca...	2

Figure 13- After data cleaning and pre-processing

For Feature Engineering using N-grams:

	N-Gram Feature Vector	Data Dimension
0	unigram	(298, 5604)
1	unigram_bigram	(298, 54038)
2	bigram	(298, 48434)
3	bigram_trigram	(298, 115329)
4	trigram	(298, 66895)

Figure 14- N-gram extraction

For building the model:

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
0	unigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.815681	0.902071
1	unigram	precision_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.860206	0.909018
2	unigram	recall_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.826720	0.912037
3	unigram_bigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.766749	0.868443
4	unigram_bigram	precision_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.812359	0.881579
5	unigram_bigram	recall_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.781481	0.884259
6	bigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.674685	0.851420
7	bigram	precision_macro	(DecisionTreeClassifier(max_depth=35, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=35, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.806748	0.855724
8	bigram	recall_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.695370	0.865741

Figure 15- Evaluation of N-gram features for Random forest classifier

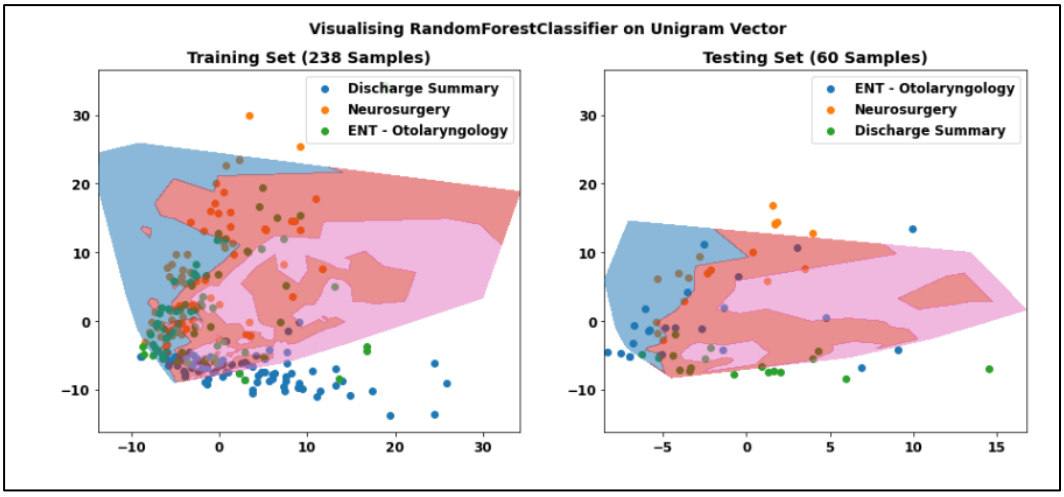

Figure 16- Visualization of Random Forest Classifier on unigram

	N-Gram Feature Vector	Data Dimension
0	unigram	(298, 974)
1	unigram_bigram	(298, 3074)
2	bigram	(298, 3334)
3	bigram_trigram	(298, 3551)
4	trigram	(298, 3139)

Figure 17- After Dimensionality reduction

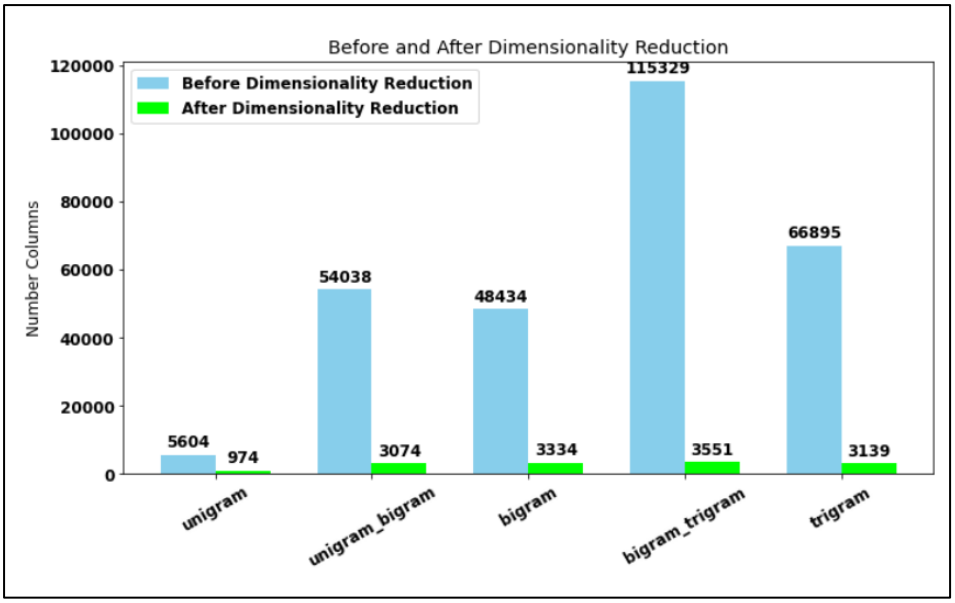

Figure 18- Comparison of dimensions for N-gram vectors

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
0	unigram	f1_macro	KNeighborsClassifier(n_neighbors=9)	0.656608	0.821654
1	unigram	precision_macro	KNeighborsClassifier(n_neighbors=9)	0.805632	0.860119
2	unigram	recall_macro	KNeighborsClassifier(n_neighbors=9)	0.668915	0.833333
3	unigram_bigram	f1_macro	KNeighborsClassifier(n_neighbors=7)	0.589224	0.790888
4	unigram_bigram	precision_macro	KNeighborsClassifier(n_neighbors=9)	0.781045	0.880952
5	unigram_bigram	recall_macro	KNeighborsClassifier(n_neighbors=7)	0.612434	0.800926
6	bigram	f1_macro	KNeighborsClassifier(n_neighbors=17)	0.182196	0.153846
7	bigram	precision_macro	KNeighborsClassifier(n_neighbors=17)	0.180259	0.100000
8	bigram	recall_macro	KNeighborsClassifier(n_neighbors=17)	0.340741	0.333333
9	bigram_trigram	f1_macro	KNeighborsClassifier(n_neighbors=17)	0.182196	0.153846
10	bigram_trigram	precision_macro	KNeighborsClassifier(n_neighbors=17)	0.180259	0.100000
11	bigram_trigram	recall_macro	KNeighborsClassifier(n_neighbors=17)	0.340741	0.333333
12	trigram	f1_macro	KNeighborsClassifier(n_neighbors=17)	0.167877	0.153846
13	trigram	precision_macro	KNeighborsClassifier(n_neighbors=17)	0.112802	0.100000
14	trigram	recall_macro	KNeighborsClassifier(n_neighbors=17)	0.333333	0.333333

Figure 19- Evaluation of N-gram features for KNeighbours classifier

```
best_vector, best_clf =  get_best_vector_clf(knn_result)
vis_classification(vector_type = best_vector, estimator = best_clf)
```

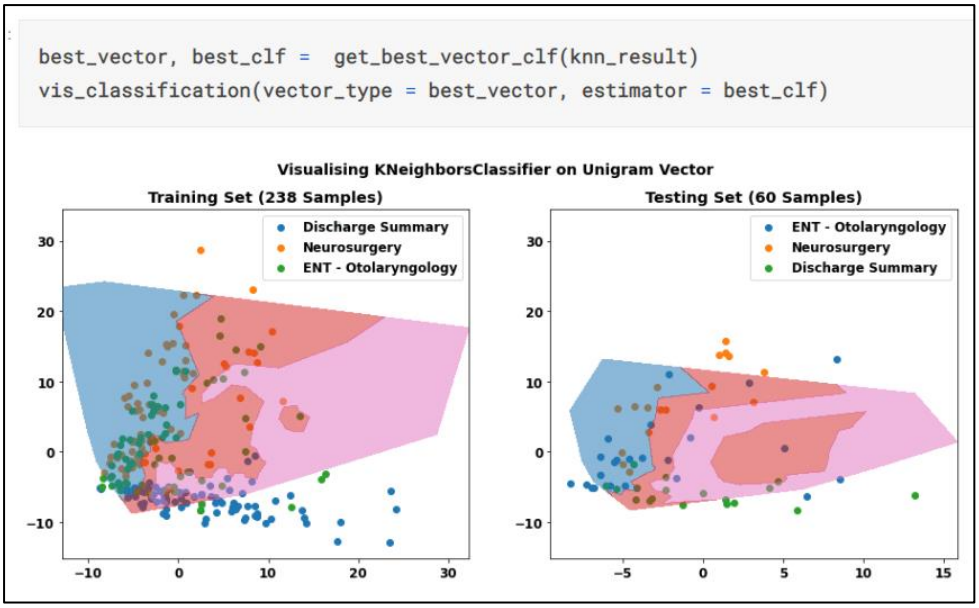

Figure 20- Visualization of KNeighbours on Unigram

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
0	unigram	f1_macro	DecisionTreeClassifier(max_depth=35, min_samples_split=5, random_state=8888)	0.846548	0.871630
1	unigram	precision_macro	DecisionTreeClassifier(min_samples_split=4, random_state=8888)	0.860370	0.881297
2	unigram	recall_macro	DecisionTreeClassifier(max_depth=8, min_samples_split=4, random_state=8888)	0.845132	0.875000
3	unigram_bigram	f1_macro	DecisionTreeClassifier(max_depth=6, min_samples_split=5, random_state=8888)	0.841640	0.870052
4	unigram_bigram	precision_macro	DecisionTreeClassifier(max_depth=6, min_samples_split=5, random_state=8888)	0.852379	0.880764
5	unigram_bigram	recall_macro	DecisionTreeClassifier(max_depth=6, min_samples_split=5, random_state=8888)	0.842725	0.870370
6	bigram	f1_macro	DecisionTreeClassifier(min_samples_split=4, random_state=8888)	0.680982	0.868120
7	bigram	precision_macro	DecisionTreeClassifier(max_depth=6, min_samples_split=5, random_state=8888)	0.797102	0.831551
8	bigram	recall_macro	DecisionTreeClassifier(max_depth=7, random_state=8888)	0.701323	0.810185
9	bigram_trigram	f1_macro	DecisionTreeClassifier(max_depth=30, min_samples_split=3, random_state=8888)	0.723130	0.869710
10	bigram_trigram	precision_macro	DecisionTreeClassifier(max_depth=7, min_samples_split=3, random_state=8888)	0.816567	0.838235
11	bigram_trigram	recall_macro	DecisionTreeClassifier(max_depth=30, min_samples_split=3, random_state=8888)	0.731481	0.865741
12	trigram	f1_macro	DecisionTreeClassifier(max_depth=30, min_samples_split=5, random_state=8888)	0.619310	0.722690
13	trigram	precision_macro	DecisionTreeClassifier(max_depth=8, random_state=8888)	0.714302	0.602083
14	trigram	recall_macro	DecisionTreeClassifier(max_depth=30, min_samples_split=5, random_state=8888)	0.648757	0.731481

Figure 21- Evaluation of N-gram features for Decision Tree classifier

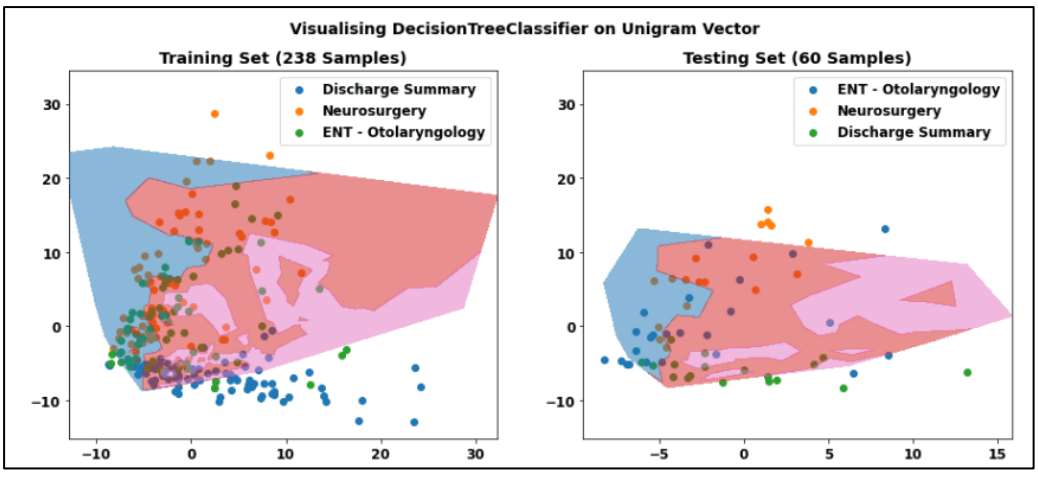

Figure 22- Visualization of Decision Tree Classifier on unigram

Training:

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
16	unigram	precision_macro	DecisionTreeClassifier(min_samples_split=4, random_state=8888)	0.86037	0.881297
17	unigram	recall_macro	DecisionTreeClassifier(max_depth=8, min_samples_split=4, random_state=8888)	0.845132	0.875
15	unigram	f1_macro	DecisionTreeClassifier(max_depth=35, min_samples_split=5, random_state=8888)	0.846548	0.87163

Figure 23- Best Classifier and Feature vector

Testing:

Evaluate on Each Class Labels

In [57]:
```
X_train, X_test, y_train, y_test = train_test_split(dataframes[best_vector], df_target, test_size=0.2, \
                                                    random_state=random_state_number)
clf = best_clf.fit(X_train, y_train)
y_test_pred= clf.predict(X_test)
target_names = ['Discharge Summary', 'ENT', 'Neurosurgery']
print(classification_report(y_test,y_test_pred,target_names=target_names))
```

```
                   precision    recall  f1-score   support

Discharge Summary       1.00      0.89      0.94        18
              ENT       0.90      0.79      0.84        24
     Neurosurgery       0.74      0.94      0.83        18

         accuracy                           0.87        60
        macro avg       0.88      0.88      0.87        60
     weighted avg       0.88      0.87      0.87        60
```

Figure 24- Evaluation

Sample Prediction:

In [60]:
```
sample_predict = pd.DataFrame({'Actual Y Test': le.inverse_transform(y_test),'Best Prediction':le.inverse_transform(y_tes
sample_predict
```

Out[60]:

	Actual Y Test	Best Prediction
0	ENT - Otolaryngology	ENT - Otolaryngology
1	ENT - Otolaryngology	ENT - Otolaryngology
2	ENT - Otolaryngology	ENT - Otolaryngology
3	ENT - Otolaryngology	Neurosurgery
4	Neurosurgery	Neurosurgery
5	Neurosurgery	Neurosurgery
6	ENT - Otolaryngology	ENT - Otolaryngology
7	Discharge Summary	Discharge Summary
8	Neurosurgery	Neurosurgery
9	ENT - Otolaryngology	ENT - Otolaryngology
10	Discharge Summary	Discharge Summary
11	Neurosurgery	Neurosurgery
12	Neurosurgery	Neurosurgery
13	ENT - Otolaryngology	ENT - Otolaryngology
14	ENT - Otolaryngology	ENT - Otolaryngology

Figure 25- Prediction results

7. RESULTS

The categorization performance was assessed using the macro F1 metric score. On the testing set, the best score obtained was over 0.8 macro F1 using tuned Random Forest and unigram feature vectors. Unigram-bigram vector also showed some good results with Decision tree classifier. Overall if the numbers increase for the N-gram vectors the classification results will have poor scores. Dimensionality Reduction is effective for such large text data.

8. CONCLUSION

In this project, I used novel NLP approaches to classify medical transcriptions into medical specialities using text data. Medical text classification is very useful in health sector. It promotes effective patient-care and better analysis for improved outcome. Natural language processing, Deep learning and Machine learning algorithms are advance techniques which should be used highly to develop clinical automatic text algorithms not only for classification but also for other problem statements.

9. FUTURE RESEARCH

Classification of medical unstructured data such as X-ray, ultrasounds, socioeconomic data, Electronic Health records (EHR), other medical images, etc have a excellent future scope which is also very useful for healthcare professionals.

10. REFERENCES

1. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text Classification Algorithms: A Survey. Information. 2019; 10(4):150. https://doi.org/10.3390/info10040150
2. M.D., M.S. in CS YH. Common Machine Learning and Deep Learning Methods for Clinical Text Classification. Medium. Published March 27, 2022. Accessed December 10, 2022. https://towardsdatascience.com/common-machine-learning-and-deep-learning-methods-for-clinical-text-classification-188473477a32
3. Yao, L., Mao, C. & Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak 19 (Suppl 3), 71 (2019). https://doi.org/10.1186/s12911-019-0781-4
4. Step 2.5: Choose a Model | Machine Learning | Google Developers. Google Developers. Accessed December 11, 2022. https://developers.google.com/machine-learning/guides/text-classification/step-2-5
5. Medical Transcriptions. Medical Transcriptions | Kaggle. Accessed December 11, 2022. /datasets/tboyle10/medicaltranscriptions
6. B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, and R. A. Dudley, "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," Journal of the American Medical Informatics Association, vol. 21, no. 5, pp. 871–875, 2014, doi: 10.1136/amiajnl-2014-002694.