Analysis of Clinical Language Models and Entity Recognizers on Medical Document Classification Task

By Peter Li, Saicharan Thirandas, Hajera Siddiqui

Abstract

Medical transcriptions provide a detailed description of a patient's condition, often forming a concise summary of a physician's observations during an examination. Accurately classifying the medical specialty (e.g. surgery, bariatrics, or neurology) based on the text transcriptions can be crucial in developing effective treatment plans and diagnosis. We use and evaluate medical named entity recognition (NER) models and pre-trained language models on publicly available datasets to provide valuable insights into the performance, accuracy, and characteristics of medical classification tasks. We will perform two types of classification tasks - medical speciality classification and medical document type classification.

1 Introduction

Natural Language Processing (NLP) has revolutionized many industries, including the healthcare sector. In healthcare, a vast amount of data is available, but much of it is unstructured. NLP has emerged as an essential tool in processing this data to provide insights and suggestions for patient care. One of the primary applications of NLP in healthcare is information extraction from clinical notes, which helps in organizing data and extracting meaningful information. Moreover, NLP can also aid in medical document classification, helping physicians identify the specialty, problem, treatment, and other critical patient history information. This makes it easier for doctors to classify patient information and recommend a suitable treatment plan from previous treatments.

However, medical text data is quite different from regular text data, and there are several challenges associated with processing it. One of the most significant challenges is variable formatting semantics, where the same word or phrase can have different meanings depending on its context. Additionally, there is no proper sentence structure, which makes it difficult to follow grammar rules, and medical jargon and abbreviations can be challenging to understand. Lastly, medical transcriptions can have misspellings and missing function punctuation, making it challenging to process the data effectively. [5]

Despite these challenges, the use of NLP in healthcare has numerous benefits, including improved accuracy, reduced manual effort, and faster data processing. This paper aims to explore the advances and challenges of NLP in healthcare, with a particular focus on medical document classification and the unique challenges of processing medical text data. By understanding these issues, we can work towards improving the efficiency and accuracy of NLP applications in healthcare, leading to better patient care and outcomes.

2 Background/Related Work

Recent advancements in natural language processing (NLP) techniques have greatly improved medical document classification, benefiting the healthcare community. Named Entity Recognition (NER) models have been trained on large medical datasets, such as i2b2 and MIMIC, to extract relevant entities such as anatomy, problems, diseases, cures, and chemicals. Medical language models and medical BERTS, such as BioBERT and ClinicalBERT, have also contributed to improving medical document classification by understanding and producing medical texts with high accuracy and efficiency. ClinicalBERT was developed to address the differences between biomedical and clinical texts, and it has been shown to outperform BioBERT in clinical NLP tasks.

In conclusion, the advancements in NLP techniques, such as NER models and medical BERTS, have greatly improved medical document classification, making it easier for healthcare providers to extract valuable information from large volumes of medical texts. The development of specialized models, such as ClinicalBERT, has further improved the accuracy of clinical NLP tasks. These advancements are expected to continue to have a significant impact on the healthcare industry.

3 Data

3.1 Initial overview of dataset

The dataset was retrieved by scraping data from mtsamples.com. This dataset contains short descriptions of the medical transcription, the medical specialty classification of the transcriptions, the title of the medical transcription, sample medical transcriptions, and relevant keywords from the medical transcription. It consists of almost 4000 rows that each contain unique medical transcriptions. The labels are not distributed evenly throughout the dataset. For example, the number of surgery labeled transcriptions amount to 1103; while the number of hospice/palliative care specialities in the dataset total to 6 (Figure 1).

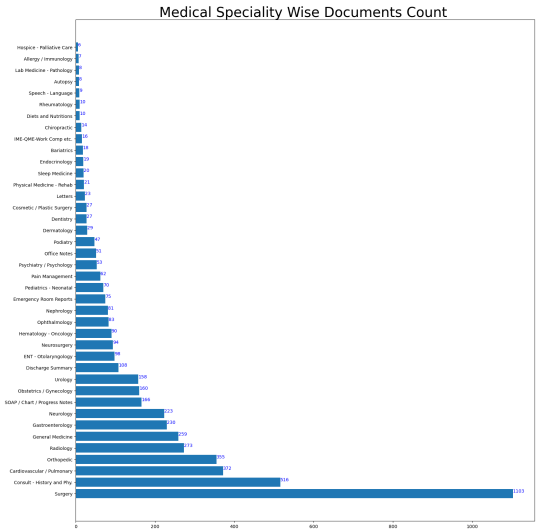

Figure 1: Skew Distribution of Labeled Transcription Specialities

Due to this obvious skew in specialty representation in the dataset, all specialities with less than 50 records in the dataset are triaged.

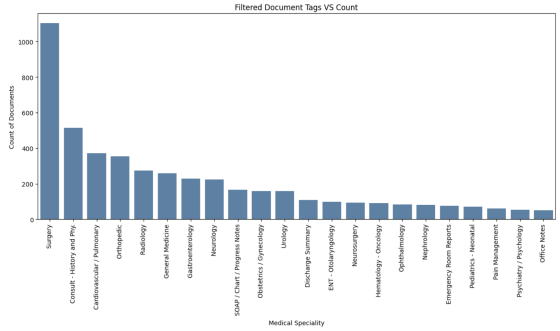

Figure 2: A Bar Plot with Categories Consisting of 50 or More Records

3.2 Why is this dataset challenging?

In this dataset, classification of certain categories in a dataset as medical specialties can be challenging. For instance, categories like 'Discharge Summary' and 'Emergency Room Reports' are not necessarily medical specialties but can contain information related to a wide range of specialties, diagnoses, or procedures. This overlap can make it difficult to classify them accurately as they can belong to multiple specialties.

Moreover, the 'Surgery' category is also problematic as it is a parent medical specialty that encompasses multiple types of specialties. For example, the 'Cardiovascular/Pulmonary' specialty has 30 records overlapping with 'Surgery', while the 'Orthopedic' specialty has 62 records overlapping with 'Surgery'. As a result, it is difficult to classify a document under this category without further information on the specific type of surgery being performed. These issues with the class labels make it challenging to accurately classify medical documents, and more advanced techniques such as natural language processing and machine learning may be needed to accurately categorize them.

The TSNE plot (Figure 3) illustrates the distribution of records in each medical specialty (where each record corresponds to a colored dot on the graph). The sparsity of specific specialties and the dense representation of other specialities is apparent in the distribution of the plot shown.

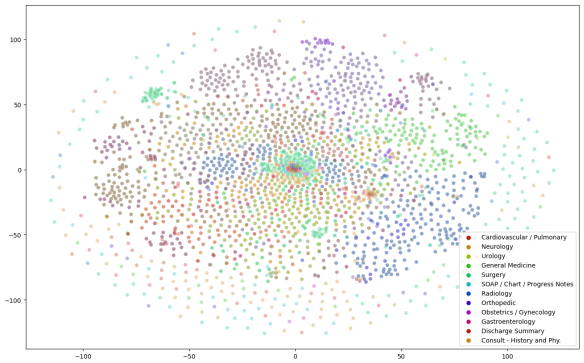

Figure 3: A TSNE Plot that Show Scatter Plot Distribution of Various Class labels**.**

3.3 Evaluate medical document label overlaps

To assert the issue of overlapping categories in medical document classification, a basic Bayes classifier was applied to the dataset, and the confusion matrix was analyzed to evaluate the accuracy of the classification. The results revealed that documents categorized as 'Surgery'

were often misclassified as cardiovascular or other specialties, while categories like 'Discharge Summary', 'Emergency Room Reports', and 'Office Notes' were also prone to being classified as various other specialties. These observations can be seen in Figure 4.

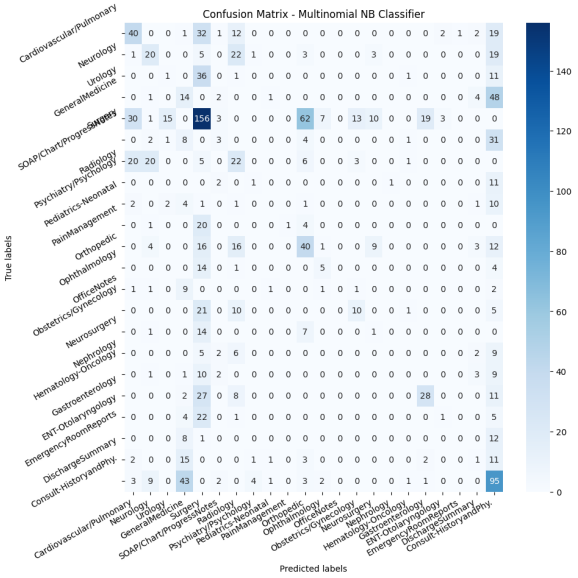

Figure 4: A Confusion Matrix that Includes all of the Categories Given in the Dataset.

3.4 Defining the Classification Tasks

To address the issues of overlapping categories in medical document classification, the dataset has been redefined into two separate classifications - one based on medical specialty and the other based on document type.

Task 1- Medical Speciality Classification :

The medical specialty task includes categories such as 'Cardiovascular/Pulmonary', 'Neurology', 'Orthopedic', and 'Gastroenterology', among others. To improve the accuracy of classification, related specialties have been merged, such as 'Nephrology' and 'Urology', and 'Neurosurgery' and 'Neurology'. In doing so, parent medical specialty classes such as 'Surgery' and 'General Medical' - which could apply to multiple classes - could be excluded for the time being, providing greater clarity and accuracy in classification.

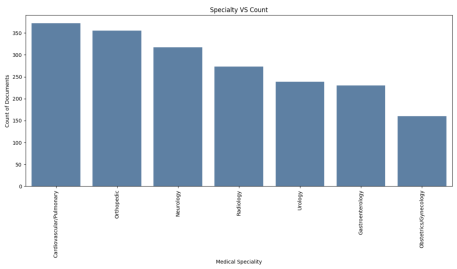

Figure 5: A Bar Plot with Medical Specialty Type Classes (Classes starting from the left to the right: Cardiovascular/Pulmonary, Orthopedic, Neurology, Radiology, Urology, Gastroenterology, Obstetrics/Gynecology)

Task 2-Medical Document Type Classification :

Similarly, the medical document type task includes categories such as 'SOAP/Chart/ProgressNotes', 'OfficeNotes', 'EmergencyRoomReports', and 'DischargeSummary'. There is a significant skew in the distribution of these categories, which can be addressed during model training. By separating the dataset into these two classifications, greater clarity and accuracy can be achieved in medical document classification.

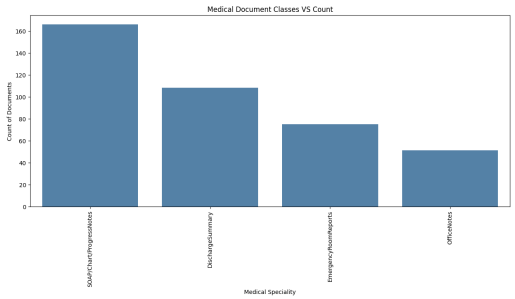

Figure 6: A Bar Plot with Medical Document Type Classes (Classes starting from the left to the right: SOAP/Chart/ProgressNotes, DischargeSummary, EmergencyRoomReports, OfficeNotes)

4 Methods

We applied both TF-IDF vectorizer with logistic regression and BERT-based models for the two medical classification tasks. For medical specialty classification, we used specific medical-related entities such as anatomy or clinical entities for TD-IDF vectorization, while for medical document type classification, we did not use medical entities as they do not play a role in identifying the document type. Instead, we focused on the transcription text itself, as different types of documents such as emergency room reports, office notes, and discharge summaries can belong to a wide variety of medical specialties and diseases. Therefore, we used BERT-based models for this task as they can capture the contextual meaning and semantic relationships between words in the text, which can be useful for accurately classifying the document type.

4.1 Feature extraction using Entity recognizers

To extract important entity information from medical transcription data, we utilized medical entity recognizer models. We used two models to extract anatomy entities and clinical entities - En_ner_bionlp13cg_md and En_ner_bc5cdr_md, respectively. [1]

Anatomy Entity Recognizer:
En_ner_bionlp13cg_md was trained on the BIONLP13CG corpus, which was part of the BioNLP Shared Task in 2013. [3] The BioNLP-ST aimed to improve information extraction for biology-based text mining, specifically focusing on six extraction tasks, including Cancer Genetics (CG). The objective of the CG task was to enhance the automatic extraction of information related to biological processes involved in cancer development and progression. [4] En_ner_bionlp13cg_md recognizes several attributes, including Anatomical_system, Cancer, Cell, Cellular_component, Organ, Organism, etc. [1]

Clinical Entity Recognizer:
On the other hand, En_ner_bc5cdr_md was trained on the BC5CDR corpus, which includes 1,500 PubMed articles with 4,409 annotated chemicals, 5,818 diseases, and 3,116 chemical-disease interactions. [7] This model specifically recognizes the attributes of Disease and Chemical. [1]

4.2 Models - TDIDF + LR Classifiers

To enhance the performance of the logistic regression model utilizing the TF-IDF vectorizer for medical specialty classification, we employed a multi-step approach. Firstly, we preprocessed the transcription data through the use of medical named entity recognition libraries, allowing us to extract relevant entities. We then trained three distinct models for medical specialty classification, one focusing solely on anatomy entities, another utilizing clinical entities, and a third using a combination of both.
Due to the high dimensionality of the data, we utilized PCA techniques to reduce the dimensionality of the TF-IDF vectorizer. Subsequently, we applied the optimized TF-IDF vectorizer to the extracted entities from the previous step. To determine the optimal number of principal components for PCA reduction, we ran the model multiple times with varying PCA values. We also addressed data imbalance by performing minority upsampling using the SMOTE technique for the case of medical document type classification tasks. Results and Experiments of them can be seen in the next section.

4.3 Models - BERT Classifiers

We investigated the performance of various pre-trained transformer-based models for medical classification tasks. Specifically, we focused on two medical-based models, BioBERT and ClinicalBERT. BioBERT is a pre-trained biomedical language representation model based on the BERT architecture and trained on large-scale biomedical text data. On the other hand, ClinicalBERT is a pre-trained biomedical language representation model based on the BERT architecture and trained on clinical notes. To facilitate our experimentation, we first created a ClassifierModel class that can load a base model from a specific Huggingface model checkpoint. We added a classification layer and instantiated the BERT model. We also created a function outside of the class that loads the BERT model and creates a ClassifierModel instance that takes in a specific classifier architecture. This allowed us to obtain the relevant tokenizer and model. Next, we initialized the BERT base models for our research. We loaded the pre-trained BioBERT model with its tokenizer, and also loaded the pre-trained BioBERT and ClinicalBERT models with their respective tokenizers. We also saved another version of the models for the data type model. We created dictionaries for the train, validation, and test datasets for both the medical specialty and data type models.

5 Experiments & Results

5.1 - Task 1 - Medical Speciality Classification

5.1.a TF-IDF + LR Classifieron Anatomy Entities(AE) and Clinical Entities (CE) .
For the below task, we have a varied pca n_components argument which automatically determines the minimum number of components

required to retain that proportion of the total variance in the data.
In the below table TD-IDF+LR + AE , refers to TFIDF + Logistic Regression over the Anatomy entities and the rest follow the same.

Model	PCA -	Precision	Recall	F1
TD-IDF+LR + AE	0.65	0.66	0.66	0.66
TD-IDF+LR + AE	0.85	0.64	0.65	0.64
TD-IDF+LR + AE	0.95	0.63	0.64	0.63
TD-IDF+LR + CE	0.65	0.63	0.64	0.63
TD-IDF+LR + CE	0.85	0.62	0.63	0.62
TD-IDF+LR + CE	0.95	0.61	0.63	0.61
TD-IDF+LR + AE+CE	0.65	0.67	0.68	0.67
TD-IDF+LR + AE+CE	0.85	0.68	0.70	0.68
TD-IDF+LR + AE + CE	0.95	0.67	0.69	0.67

Table to report accuracy over multiple experiments of TF-IDF for medical specialty classification task

Conclusion : From the above table we can conclude that when model is applied on anatomy entities and clinical entities together it performs better than individual entities

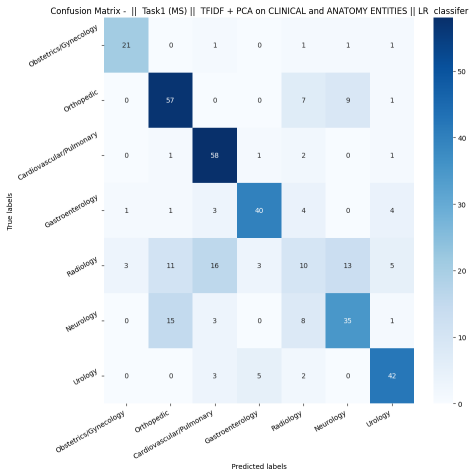

Figure 7: A Confusion Matrix that Includes Task 1 of TF-IDF and PCA on Clinical and Anatomy Entities

	precision	recall	f1-score	support
Obstetrics/Gynecology	0.84	0.84	0.84	25
Orthopedic	0.67	0.77	0.72	74
Cardiovascular/Pulmonary	0.69	0.92	0.79	63
Gastroenterology	0.82	0.75	0.78	53
Radiology	0.29	0.16	0.21	61
Neurology	0.60	0.56	0.58	62
Urology	0.76	0.81	0.79	52
accuracy			0.67	390
macro avg	0.67	0.69	0.67	390
weighted avg	0.65	0.67	0.65	390

Figure 8: A Classification Report of Analysis TFIDF + on Clinical and Anatomy Entities

5.1.b BERT Based Models - Bert, BioBert, ClinicalBert

We have trained BERT + FNN classifiers with varying architectures (e.g. 1-4 hidden layers and differing hyperparameters), the best of which is outlined in the table. Each BERT model was fine tuned for around 30 Epochs and results of which can be seen below. We used dropout (p=0.5) and ReLU as the models activation function.

Model	Learning Rate	Batch Size	Classifier Architecture	Validation Accuracy	F1	Epochs
Base BERT	0.0001	8	Input 256 Output	0.4327	0.37	30
BioBERT	0.0001	8	Input 256 Output	0.6923	0.66	30
ClinicalBERT	0.001	8	Input 256 Output	0.6827	0.65	30

Table to report accuracy over multiple experiments of BERT ,BioBERT and ClinicalBERTs

As expected we can clearly see the significant increase in the model performance of BioBERT and ClinicalBERT compared to the Basic BERT model. Nonetheless, we also expected an improvement of the ClinicalBERT model compared to the BioBERT model. However, this did not occur. It did not perform as expected because medical speciality classification rely on anatomy related embeddings than other clinical specific embeddings.

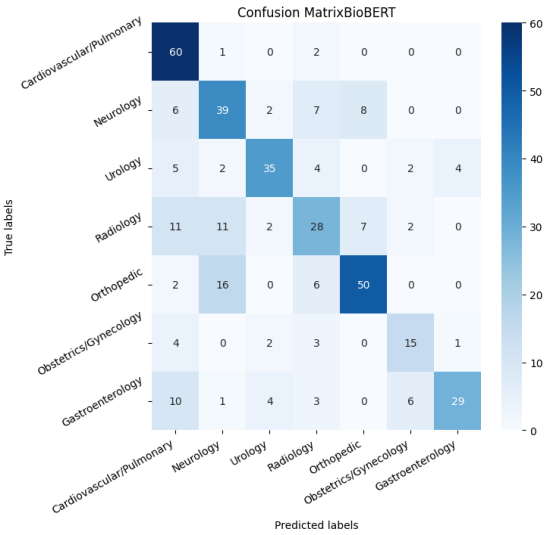

Figure 9: A Confusion Matrix of BioBERT + FNN Classifier

	precision	recall	f1-score	support
Cardiovascular/Pulmonary	0.61	0.95	0.75	63
Neurology	0.56	0.63	0.59	62
Urology	0.78	0.67	0.72	52
Radiology	0.53	0.46	0.49	61
Orthopedic	0.77	0.68	0.72	74
Obstetrics/Gynecology	0.60	0.60	0.60	25
Gastroenterology	0.85	0.55	0.67	53
accuracy			0.66	390
macro avg	0.67	0.65	0.65	390
weighted avg	0.67	0.66	0.65	390

Figure 10: A Classification Report of BioBERT + FNN Classifier

We also looked at the BERT, BioBERT, and ClinicalBERT fine tuned models. The BERT model produced a training loss of 1.6605 and a validation accuracy of 0.4038 with 30 epochs. The BioBERT model produced a training loss of 0.8086 and a validation accuracy of 0.6603 with 30 epochs. The ClinicalBERT model produced a training loss of 0.6197 and a validation accuracy of 0.6410 with 30 epochs. As we can see, BioBERT and ClinicalBERT models work much better than the BERT model. Although BioBERT and ClinicalBERT have results that are quite similar, BioBERT did perform slightly better.

5.2 Task 2 - Medical Document Type Classification

5.2.a TFIDF Classifier +LR on Transcriptions (TS)

Model	**PCA**	**Precision**	**Recall**	**F1**
TD-IDF+LR + Transcription	0.65	0.55	0.51	0.50
TD-IDF+LR + Transcription	0.85	0.57	0.52	0.52
TD-IDF+LR + Transcription +	0.95	0.59	0.58	0.56
TD-IDF+LR + Transcription + SMOTE	0.65	0.59	0.53	0.53
TD-IDF+LR + Transcription + SMOTE	0.85	0.60	0.54	0.54
TD-IDF+LR + Transcription + SMOTE	0.95	0.64	0.59	0.57

Table to report accuracy over multiple experiments of TF-IDF for medical document type classification task

From the above table we can conclude that when SMOTE is applied to overcome class imbalance issue there is significant increase in the F1 score which can be clearly seen.

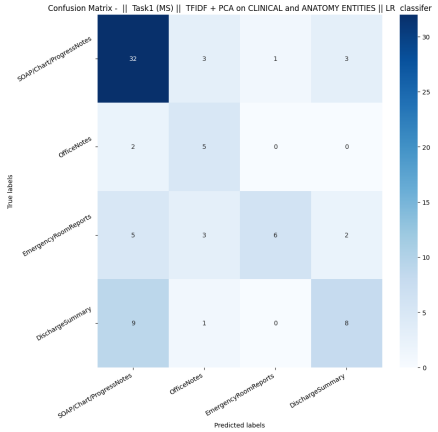

Figure 11: A Confusion Matrix of Task 2 of TF-IDF and LR on Clinical and Anatomy Entities

	precision	recall	f1-score	support
SOAP/Chart/ProgressNotes	0.67	0.82	0.74	39
OfficeNotes	0.42	0.71	0.53	7
EmergencyRoomReports	0.86	0.38	0.52	16
DischargeSummary	0.62	0.44	0.52	18
accuracy			0.64	80
macro avg	0.64	0.59	0.57	80
weighted avg	0.67	0.64	0.63	80

Figure 12: A Classification Report of Task 2 of TF-IDF and LR on Clinical and Anatomy Entities

5.1.b BERT Based Models - BERT, BioBERT

Similarly, here trained BERT + FNN classifiers which have 2 hidden layers consisting of 512 and 128 units. Each BERT model was fine tuned for around 30 Epochs and results of which can be seen below.

	Learning Rate	**Batch Size**	**Classifier Architecture**	**Validation Accuracy**	**F1 Score**	**Epochs**

Base BERT	0.0001	8	Input 256 Output	0.6406	0.50	30
BioBERT	0.0001	8	Input 256 Output	0.8281	0.73	30

Table to report accuracy over multiple experiments of BERT and BioBERT

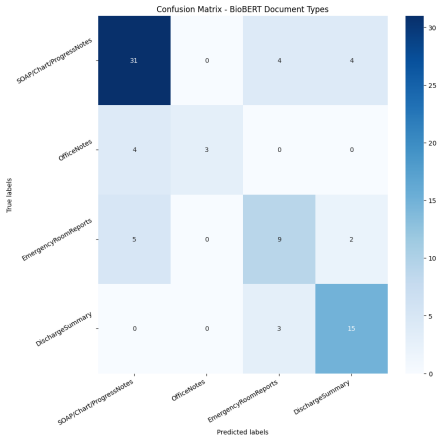

Figure 13: A Confusion Matrix of BioBERT + FNN Classifier to Classify Medical Document Types.

	precision	recall	f1-score	support
SOAP/Chart/ProgressNotes	0.78	0.79	0.78	39
OfficeNotes	1.00	0.43	0.60	7
EmergencyRoomReports	0.56	0.56	0.56	16
DischargeSummary	0.71	0.83	0.77	18
accuracy			0.73	80
macro avg	0.76	0.65	0.68	80
weighted avg	0.74	0.72	0.72	80

Figure 14: A Classification Report of BioBERT + FNN Classifier to Classify Medical Document Types.

6 Conclusion & Future works

As observed from our confusion matrices, we had a lot of overlapping between categories. There are a few orthopedic records that are getting misclassified as the neurology specialty. A reason for this could be that some orthopedic surgeons focus on nervous systems disorders that could involve the nerves or the spinal cord. Neurology also deals with the nervous system and the spinal cord.

Another takeaway from the confusion matrix is that 'Radiology' is overlapped with multiple other specialties. Reasonings for this could involve the fact that radiology involves using imaging technology to scan a certain part of the body. It makes sense why it would have a lot of overlap with categories such as 'Orthopedic','Cardiovascular/Pulmonary', 'Neurology', due to scanning the bones and joints, heart, and the brain, respectively.

In order to address the issues observed in our current approach, we can pursue several strategies in future work. These may include experimenting with different classifier network architectures on top of BERT models, increasing the number of training epochs, and evaluating their impact on accuracy. Additionally, exploring alternative BERT models such as MedBERT, GatorTron, and BioGPT could provide insights on their suitability for our specific use case. Adopting hierarchical classification approaches for all 40 initial document tags and leveraging available labels to assign multiple tags to each document could improve the overall categorization process.

7 Citations

[1] Allen. "scispacy | SpaCy models for biomedical text processing." GitHub Pages, https://allenai.github.io/scispacy/. Accessed 15 April 2023.

[2] Alsentzer, Emily, et al. "Publicly Available Clinical BERT Embeddings." https://aclanthology.org/W19-1909.pdf. Accessed 15 April 2023.

[3] BioNLP-ST 2013, https://2013.bionlp-st.org/. Accessed 15 April 2023.

[4] "Cancer Genetics (CG) task." BioNLP-ST 2013, https://2013.bionlp-st.org/tasks/cancer-genetics-cg-task. Accessed 15 April 2023.

[5] Leaman, Robert, et al. "Challenges in clinical natural language processing for automated disorder normalization." Journal of Biomedical Informatics, vol. 57, 2015, pp. 28-37. https://www.sciencedirect.com/science/article/pii/S1532046415001501. Accessed 15 April 2023.

[6] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." vol. 36, no. 4, 2020, pp. 1234–1240, https://academic.oup.com/bioinformatics/article/36/4/1234/5566506?login=false. Accessed 15 April 2023.

[7] Li, Jiao. "BC5CDR Dataset." Papers With Code, https://paperswithcode.com/dataset/bc5cdr. Accessed 15 April 2023.