HEALTHCARE TEXT ANALYTICS USING RECENT ML TECHNIQUES AND DATA CLASSIFICATION USING AWS CLOUD ML SERVICES

by

Movinuddin

A Dissertation (Thesis) Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

Computer Science

Middle Tennessee State University

October, 2023

Dissertation Committee:

Dr. Khem Poudel, *Chair*

Dr. Jorge Vargas

Dr. Jaishree Ranganathan

I dedicate this thesis to my parents and the people who have supported me throughout my education.

Thanks for everything!

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my supervisors, Dr. Khem Poudel, Dr. Jorge Vargas and Dr. Jaishree Ranganathan for their invaluable advice, continuous support, and patience during my masters study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I would also like to thank the two anonymous reviewers for their comments and suggestions. I would like to thank all the members in the department of computer science, MTSU. It is their kind help and support that have made my study and life in the US a wonderful time. Finally, I would like to express my gratitude to my parents. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

ABSTRACT

Classification of clinical texts has a significant impact on disease diagnosis, medical research and automated development of disease ontologies. Because they contain terms that describe medical concepts and terminology, the data set is quite noisy and the text in the transcriptions overlaps with the categories making clinical text difficult to classify. The clinical narrative, which provides a patient's history and evaluations as well as data for clinical decision-making, is the main form of communication in the medical field. The aim of the study is to make disease diagnoses based on medical records using ML algorithms. The proposed clinical text classification model using weak monitoring to reduce the human efforts to create labeled training data and conduct feature engineering. The primary objective is to contrast this approach with a logistic regression model to classify medical records clinical text and expect superior performance compared to the logistic regression model for an imbalanced medical transcriptions dataset. A promising intelligent data-driven health system to archive and classify healthcare records relies on the ability to extract and contextualize unstructured medical data in a form of a single easy-to-use API by leveraging Machine Learning (ML) services from Amazon cloud in a clinical workflow. AWS services such as S3, Textract, Comprehend Medical, and DynamoDB can be integrated to create a comprehensive solution for handling medical document processing, extracting medical information, performing medical text analysis, and storing the data in a structured manner.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I : INTRODUCTION

In the field of NLP, text classification is one of the most important fields and it helps in the assignment of the text documents to proper classes depending on their content. Many challenges and solutions are exhibited by the publicly available documents and its classification is mainly intended for web classification, unstructured text classification, sentiment classification, spam e-mail filtering and author identification [2]. The medical records include the doctor's examination, diagnosis procedures, treatment protocols and notification of improvement of the disease in the patient. The entire medical history, along with the prescription effect of the medicine on the patient, is also stored in the medical record. The medical literature includes the oldest and recent documents of the medical techniques used for diagnosis and treatment of a particular disease. Both information resources are very important in the field of clinical medicine. Due to the advent of information technology, a tremendous quantity of electronic medical records and literature have been found online, which provides good resources for data mining in the medical field. Text classification in the medical field is quite challenging because of two main issues: first, it has a few grammatical mistakes, and second, a lot of medical techniques are presented in the text [3].

For the past several decades, a community of researchers working at the intersection of computer science and medicine have developed strategies for information extraction and modeling of clinical text, using techniques somewhat distinct from those of the broader natural language processing (NLP) research community [4]. Their efforts have led to the development of new methods and the production of both commercial and open-source software systems for clinical text mining [5]. In recent years, technology giants like Amazon and Google have also recognized the importance of clinical text mining and joined the Fray. Amazon Comprehend Medical now comes packaged as a software add-on to Amazon Web Services, incentivizing storage of Electronic Health Record (EHR) data on Amazon's HIPAA (Health In- surance Portability and Accountability Act)-compliant cloud platform

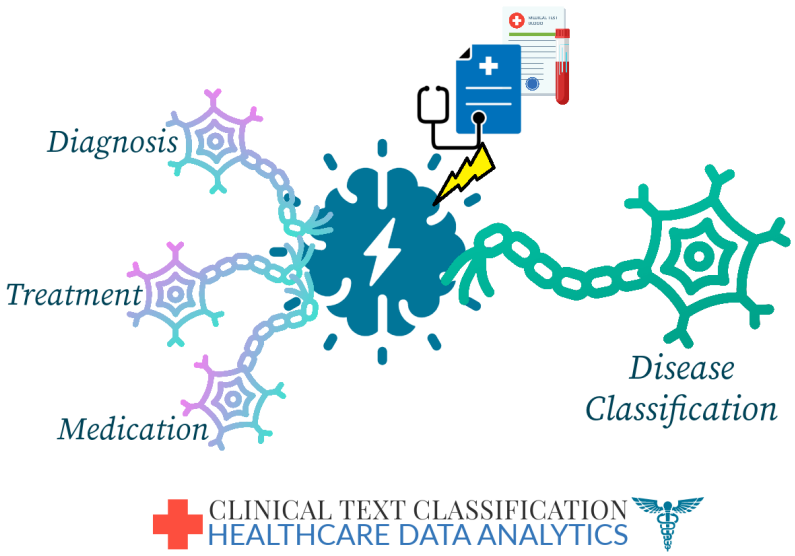

Figure 1.1: Clinical NLP [1]

by providing seam- less clinical text processing. Dedicated clinical text–processing companies such as Clinithink, Linguamatics, and Apixio have built proprietary systems of their own, promising to improve clinical trial recruitment, disease registry creation, government reporting, and billing, all through improved mining of unstructured clinical text [6]. Thus, the analysis of unstructured data with a novel solution for data sensitivity, security, quality, and accessibility using ML and NLP should be proposed. The main goal of this research is to develop ML and NLP method for recognizing unstructured clinical text data and classify it [7].

Amazon Web Services (AWS) offers a powerful suite of services that can be combined to create a solution for managing medical documents, extracting information, and performing analysis. Amazon S3 (Simple Storage Service) is AWS's object storage service, ideal for storing and retrieving data. You can upload and store your documents (PDFs, images, etc.) in S3 buckets. Amazon Textract is an AWS service for extracting text and data from scanned documents, images, and PDFs. It uses ML to analyze the documents and extract structured data. We can configure S3 buckets to trigger Textract automatically when new

documents are uploaded. Amazon Comprehend is a NLP service that can be used to perform sentiment analysis, entity recognition, key phrase extraction, and language detection. Once Textract extracts the text from documents, we can use Comprehend to analyze the extracted text for insights. This combined solution of AWS S3, Textract, and Comprehend medical can automate healthcare document processing, text extraction, and provide deeper insights into the content of documents, enabling various analytical possibilities.

CHAPTER II : BACKGROUND

Healthcare Clinical Text Classification of medical transcriptions based on various diseases has seen some of the greatest growth. NLP for Healthcare Using NLTK, spaCy, TF-IDF, and Word Embedding with non-sequence and sequence modeling like LSTM and Transformer has received critical acclaim for revealing hidden information in clinical texts. NLP has become a hot topic in AI research and applications as ML and deep learning (DL) algorithms have advanced because text, such as English sentences, is a significant type of natural language data. Text is often present in healthcare datasets like EHR. Clinical Named Entity Recognition (NER), clinical text classification, and other similar methods are among the many AI approaches to text data in the healthcare industry. Clinical text analysis is a topic of study that involves using NLP and other computational approaches to extract and evaluate information from clinical texts, such as electronic medical records (EMRs), clinical notes and other types of healthcare documents. The purpose of clinical text analysis is to find patterns and trends in clinical data and to use this knowledge to improve patient care, progress medical research and optimize healthcare systems. To extract and classify information from clinical texts, several NLP approaches such as tokenization, part-of-speech tagging, and named entity identification are used. It also entails the application of ML algorithms to find patterns and trends in data. Hospitals health clinics and offices of practicing physicians all have access to historical data on public health care [8]. Unfortunately, fewer of those data have been utilized to assist physicians or others in learning new information and comprehending their patients' health conditions in society at large. With predictions based on health care history data, expert systems and decision-making may be able to use health care history data as a knowledge base to ascertain a person's health status [3].

A Medical Records Archival application serves a crucial role in the healthcare industry. Healthcare providers are required to maintain patient records for extended periods to comply with legal and regulatory requirements. An archival application ensures secure,

long-term storage and easy retrieval of historical patient data, meeting compliance standards like HIPAA. Over time, patient records accumulate. An archival system allows for efficient organization, indexing, and retrieval of older records, enabling healthcare professionals to access patient data quickly, aiding in decision-making and patient care. Digitizing and archiving medical records reduces the need for physical storage, cutting down on paper-based storage costs, physical space requirements, and the time spent searching for and managing paper documents. Archived records provide valuable data for research, trend analysis, and understanding patient histories over time [9]. This aids in identifying patterns, improving treatments, and contributing to medical research.

CHAPTER III : HEALTHCARE TEXT ANALYTICS USING RECENT ML TECHNIQUES

Through the extraction of insightful information from massive amounts of unstructured clinical and medical text data, healthcare text analytics, enabled by modern ML techniques, has the potential to completely transform the healthcare sector. Various uses for these insights include disease diagnosis, therapy suggestions, patient management, and medical research [8]. Healthcare text analytics can leverage recent ML techniques as mentioned below:

- EHR Analysis: This can be analyzed by ML models to extract structured data such patient demographics, medical history, and test results. Unstructured clinical notes and narratives can be mined for information using NLP techniques. Named Entity Recognition (NER) models may recognize and classify items such as illnesses, signs, treatments, and procedures described in clinical writing.

- Clinical Decision Support: Medical personnel can access real-time decision support from ML algorithms by analyzing clinical text data. For instance, using symptoms, a person's medical history, and the results of tests, ML models can help diagnose diseases. By taking into account patient-specific data, recommended treatments, and the most recent medical research, machine learning models can suggest individualized treatment strategies.

- Pharmacovigilance: By examining adverse event reports, social media data, and medical literature, ML can improve medication safety monitoring. Potential safety issues and new trends can be found using sentiment analysis and topic modeling. Drug interactions, contraindications, and adverse effects can all be automatically detected using ML models.

- Healthcare Fraud Detection: Healthcare providers and payers can reduce costs by identifying fraudulent claims and activity in insurance data using ML approaches, such as anomaly detection and predictive modeling.
- Disease Outbreak Prediction: To spot illness outbreaks early on, ML models can examine text data from the healthcare industry as well as news articles and social media posts. In order to prevent the spread of diseases, this can assist public health organizations.
- Patient Engagement and Chatbots: ML-powered chatbots and virtual assistants can answer patient queries, schedule appointments, and provide health-related recommendations based on natural language understanding and generation.
- Clinical Research: Based on NLP and generation, ML-powered chatbots and virtual assistants may book appointments, respond to patient questions, and provide recommendations regarding their health.
- Radiology and Medical Imaging: To help radiologists find anomalies and diseases like cancer and fractures, ML models may examine radiology reports and medical images.

The accuracy and performance of NLP jobs in healthcare text analytics have considerably improved recently because to deep learning developments including transformer-based models like BERT and GPT. Additionally, ML models can be improved upon using data particular to the healthcare industry thanks to transfer learning approaches, which can increase their efficacy. However, while deploying healthcare text analytics solutions, it's crucial to take regulatory compliance and data privacy into consideration because healthcare data is extremely sensitive and governed by tight standards like HIPAA in the United States. To guarantee patient data security and compliance with healthcare laws, proper data anonymization and encryption are essential.

Multi-class classification of medical speciality

Performing multi-class classification in the context of healthcare specialties can be achieved using logistic regression with Scikit-Learn. We have a dataset where each sample is associated with a particular healthcare specialty (such as cardiology, oncology, neurology, etc.) and various features related to patients or medical records.

Scikit-Learn

Scikit-Learn, a popular machine learning library in Python, provides various models and tools for implementing logistic regression for multi-class classification.

Logistic Regression

In Python using scikit-learn, we can build a ML model for multi-class classification using logistic regression.Texts are preprocessed using spaCy or scispaCy for tokenization, lemmatization, and stop word removal.TF-IDF vectorization is used to convert the processed text data into numerical features. A multi-class logistic regression model is trained and evaluated using the LogisticRegression class from scikit-learn. Finally, the model is used to make predictions on new text data.

SMOTE

SMOTE stands for Synthetic Minority Over-sampling Technique, a method used in dealing with imbalanced datasets, especially in the context of machine learning classification tasks.

Imbalanced datasets, where one class is significantly more prevalent than another, can lead to biased models that favor the majority class. SMOTE is a resampling technique that addresses this issue by oversampling the minority class, creating synthetic samples rather than replicating existing ones [10].

SMOTE works by generating synthetic examples from the minority class. It does this by selecting a sample from the minority class and finding its k-nearest neighbors. It then creates synthetic examples along the line segments joining these k-nearest neighbors [11].

Key Steps in the SMOTE Algorithm:

- Selecting a Sample: Choose a sample from the minority class.
- Finding Nearest Neighbors: Identify its k-nearest neighbors in the feature space.
- Creating Synthetic Samples: Generate synthetic samples by interpolating between the chosen sample and its neighbors.

SMOTE is an effective technique in handling imbalanced datasets by creating synthetic samples, thus improving the robustness and accuracy of models, particularly in situations where the class imbalance is significant [12].

Assumptions

- Clinical Notes recorded in unstructured format.
- Clinical Notes contain vast amount of information.
- If there are certain medical specialty samples are in minotity as compared to other samples, it is imbalanced dataset.

Delimitations

- To develop ML-based Health Care text analysis models for disease classification.
- To classify medical specialty based on clinical transcriptions using Logistic Regression Classifier and SMOTE Algorithms.

CHAPTER IV : DATA CLASSIFICATION USING AWS CLOUD ML SERVICES

An IT solution to archive medical records and analyze its text to auto-classify it using AWS Cloud ML Services.

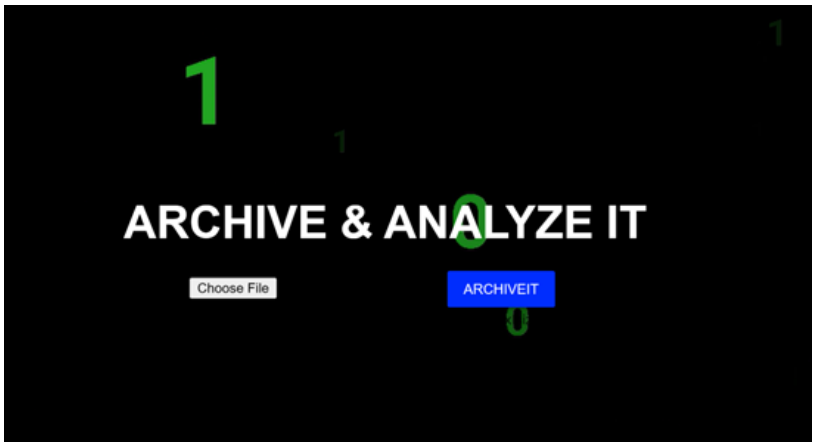

Figure 4.1: Medical Archive and Analyze IT Application

According to the cloud computing paradigm, data is permanently kept on internet servers and temporarily cached on clients, such as desktops, media centers, table computers, notebooks, wall computers, handheld devices, sensors, and monitors.

There are several service models and deployment models in cloud computing as mentioned below [13].

Service Models:

- Software-as-a-Service (SaaS): Virtual machines, storage, and networking are among the virtualized computing resources that users can rent. Both the operating system and the apps are under their control.

- Platform-as-a-Service (PaaS): PaaS offers a platform with the necessary infrastructure, tools, and services to create, launch, and manage applications. Users concen-

trate on creating applications while the cloud service provider manages the infrastructure.

- Infrastructure-as-a-Service (IaaS): Virtualized computing resources, such as virtual machines, storage, and networking, are made available to customers through IaaS. Users don't have to spend money on actual hardware because they can manage and configure these resources. IaaS providers like Amazon Web Services (AWS) and Microsoft Azure are two examples.

Cloud deployment models include:

- Public Cloud: Services are offered to the general public and are hosted and run by outside cloud service providers. Offerings for the public cloud are frequently affordable and scalable. AWS, Azure, and Google Cloud Platform (GCP) are significant public cloud service providers.

- Private Cloud: A private cloud can be hosted on-site or by a third-party provider and is exclusive to a single company. Private clouds provide more flexibility, security, and control possibilities.

- Hybrid Cloud: The components of both public and private clouds are combined in a hybrid cloud, enabling data and applications to be transferred across them. It provides adaptability and workload optimization.

Below are the Cloud Computing Benefits

- Accessibility anywhere, with any device

- Ability to get rid of most or all hardware and software

- Centralized data security

- Higher performance and availability

- Quick application deployment
- Instant business insights
- Business continuity
- Price-performance and cost savings
- Virtualized computing
- Cloud computing is greener

ECM Cloud Using AWS Services

Solutions for Enterprise Content Management that are hosted and provided using cloud computing infrastructure are referred to as ECM Cloud. An organization can capture, manage, store, preserve, and deliver material and documents linked to their business processes using an ECM, which is a collection of strategies, processes, and tools.

It has following benefits:

- Faster configuration
- User Improved user experience and accessibility
- Data location
- Reduced Cost
- Security and Protection
- Accessibility and Productivity
- Automatic updates, reduced costs and efforts

Here's a solution that incorporates these services to create a medical cloud to store and analyze healthcare records.

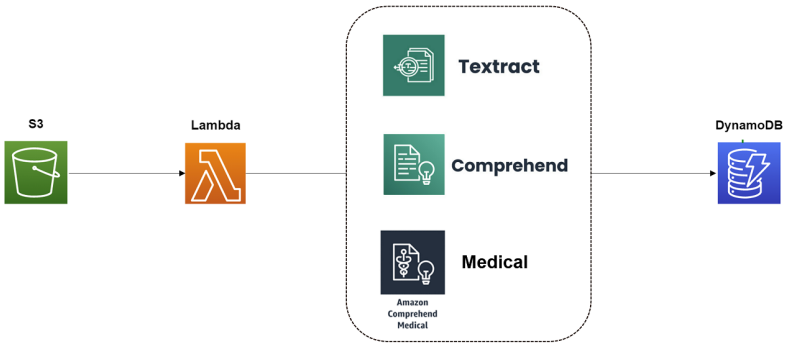

Figure 4.2: AWS Cloud Block Diagram

- Amazon S3 (Simple Storage Service): Use S3 for storing medical documents like patient records, medical reports, images, and other related data.
- Amazon Textract: Textract is used to extract text and data from medical documents, such as medical reports, patient history, etc. It automatically identifies and extracts information from various document types [14].
- Amazon Comprehend Medical: Comprehend Medical is an NLP service designed specifically for extracting medical information and identifying entities like medical conditions, medications, dosages, treatment details, etc., from unstructured medical text [15].
- DynamoDB: DynamoDB is a NoSQL database service that provides high-performance, scalable storage. It can be used to store structured medical information extracted from documents.

Below are the steps in our workflow.

- Set up Amazon S3: Create S3 buckets to store medical documents securely. Configure access permissions and encryption as per compliance requirements.

- Use Amazon Textract: Configure S3 bucket event notifications to trigger Textract when new medical documents are uploaded. Textract will analyze the documents, extract text and structured data related to patients, treatments, medications, etc.

- Use Amazon Comprehend Medical: Pass the extracted text from Textract to Comprehend Medical for medical entity extraction and analysis. Utilize Comprehend Medical's APIs to extract medical information like medical conditions, medications, dosages, etc.

- Store Data in DynamoDB: Create a DynamoDB table to store the structured medical information extracted from documents. Define the schema based on the entities extracted by Comprehend Medical (e.g., patient ID, medical condition, medication, dosage, date, etc.).

- Application Integration: Integrate the data stored in DynamoDB into your applications or medical systems for retrieval, analytics, or further processing.

Storage S3

One of the most well-known and often used cloud storage services provided by AWS is Amazon S3 (Simple Storage Service). For a variety of applications, from straightforward data storage to intricate big data analytics and content distribution, it offers scalable, resilient, secure, and highly available object storage.

Key features and concepts of Amazon S3 include:

- Objects: Data is kept in Amazon S3 as objects, each of which includes the data itself, a special key (such a filename), and metadata. The size of an object might vary, from a few bytes to many terabytes.

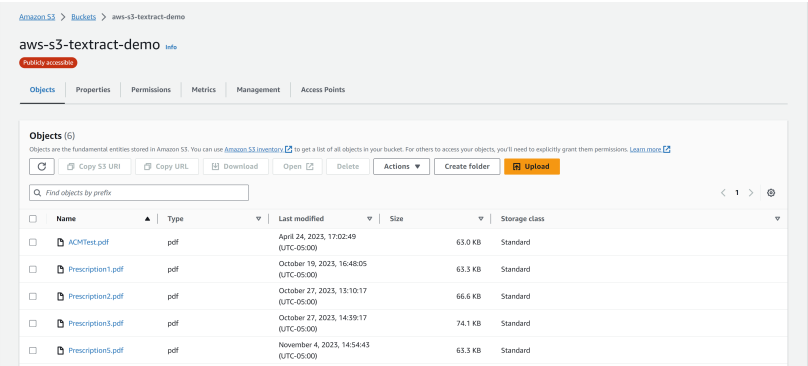

Figure 4.3: AWS S3 Storage

- Buckets: In Amazon S3, data is arranged into units called buckets. To manage your data, you can establish many buckets in your AWS account. Each bucket has a name that is globally unique.

- Data Durability and Availability: With a data durability of 99.999999999 percent [16] over a given year, Amazon S3 is made for high durability. Additionally, it offers high availability, guaranteeing that your data is available when required.

- Data Access Control: To limit who may access your data and what they can do with it, S3 provides a comprehensive set of access control tools, such as bucket policies and Access Control Lists (ACLs). You have the option of making an item public or limiting access to certain AWS users or roles.

- Data Encryption: Both in-transit and at-rest encryption are supported by S3. You can encrypt data at rest using server-side encryption (SSE) and data in transit using SSL/TLS. The SSE-S3, SSE-KMS (Key Management Service), and SSE-C (Customer-Provided Keys) options are available.

- Versioning: You may retain several versions of any object in a bucket by turning on versioning in S3. For data backup and recovery scenarios, this can be helpful.

- Lifecycle Policies: To automatically move items to other storage classes (such as from Standard to Glacier for archiving) or remove things that are no longer required, you can specify lifecycle policies.

- Cross-Region Replication: By enabling object replication across many AWS regions, S3 delivers disaster recovery capabilities and low-latency access to data across numerous locations.

- Event Notifications: You can set up event alerts to launch specified actions (such calling AWS Lambda functions) in response to certain S3 bucket events, including the addition of new objects.

- Integration: Amazon S3 is a flexible storage solution for a variety of use cases, including data lakes, backup and restore, website hosting, and content distribution via Amazon CloudFront. Amazon S3 is widely integrated with other AWS services and third-party tools.

- Organizations of all sizes frequently utilize Amazon S3 as the foundation for cloud-native apps because it offers a highly dependable and affordable solution to store and manage data in the cloud. Because of its scalability and versatility, it is a crucial part of many AWS-based solutions.

Lambda

AWS offers the serverless compute service known as Lambda. You can use it to execute code in response to events without having to set up or manage servers. Lambda is a potent tool for many use cases, including data processing, real-time file processing, automation, and more. With Lambda, you can create serverless applications that automatically scale in response to incoming requests or events [17].

Below are some key features and concepts associated with AWS Lambda:

- Event-Driven Programming: Events such as changes to data in an S3 bucket, incoming HTTP requests made through Amazon API Gateway, or events produced by other AWS services like AWS SNS, AWS SQS, or AWS CloudWatch trigger the execution of AWS Lambda functions, which are event-driven.

- No Server Management: You don't have to be concerned about infrastructure management, server provisioning, scaling, or patching when using Lambda. You can concentrate entirely on your code because AWS will take care of all the supporting infrastructure.

- Supported Languages: Programming languages supported by Lambda include Node.js, Python, Java, C#, Ruby, and Go. The language that best fits your application can be chosen.

- Stateless Functions: Because they are stateless by design, lambda functions don't store any state data between calls. To store and maintain state, you can, if necessary, leverage other AWS services like AWS DynamoDB or AWS S3.

- Scaling: In reaction to the volume of incoming events or the configured concurrency level, lambda functions can scale automatically. Your application can handle different workloads without manual assistance thanks to this scalability.

- Pay-as-You-Go Pricing: The pricing structure for AWS Lambda is pay-as-you-go. You are charged based on the quantity of requests and the length of time that your functions are executed; a free tier is offered for a certain amount of usage.

- Execution Environment: Your code can run in a containerized environment thanks to AWS Lambda. It takes care of managing your dependencies, runtime, and infrastructure.

- Triggers: Events called triggers call Lambda functions. These could be requests over the API Gateway, modifications to AWS resources (such the creation of S3 objects), unique events, or scheduled operations.

- VPC Integration: As long as security and isolation are maintained, Lambda functions can access resources in your private network by running inside an Amazon Virtual Private Cloud (VPC).

- Logging and Monitoring: For logging and monitoring purposes, AWS Lambda interfaces with AWS CloudWatch, making it simple to trace the execution of your functions and troubleshoot problems.

- Security: IAM (Identity and Access Management) roles can be set for Lambda functions to manage the access and permissions they need to AWS resources.

A useful tool for creating serverless apps that are both highly scalable and reasonably priced is AWS Lambda. It is a popular choice for a variety of use cases, from straightforward automated chores to intricate microservice architectures, because it frees developers up to concentrate on building code and logic rather than dealing with the administrative burden of managing infrastructure.

Comprehend

AWS offers a NLP service called Amazon Comprehend [18]. It is intended to assist developers and businesses in deriving insights from unstructured text data, like those found in documents, news stories, customer reviews, and social media posts. Various NLP activities are carried out by Amazon Comprehend using ML models to glean useful information from text sources.

Below are the key features and capabilities of Amazon Comprehend

- Text Classification: Amazon Comprehend is helpful for content tagging and managing massive datasets since it can automatically classify documents or text into certain categories or specified classes.
- Sentiment Analysis: In order to help businesses understand client thoughts, feelings, and trends regarding their goods and services, it can identify the sentiment (positive, negative, or neutral) expressed in text.
- Entity Recognition: From text, Comprehend can recognize and extract entities like individuals, companies, places, dates, and more. This is helpful for data enrichment and information extraction.
- Keyphrase Extraction: In order to help readers grasp the primary issues or ideas within a document, it recognizes and extracts significant phrases or topics from text.
- Language Detection: For multi-language content analysis, Amazon Comprehend can automatically identify the language of text documents.
- Syntax Analysis: It has the ability to parse syntax trees and tag the parts of speech in sentences as well as assess their grammatical structure. For further in-depth language study, this can be helpful.
- Custom Entity Recognition: To recognize domain-specific entities in text data, users can train their own bespoke entity recognition models.
- Multi-Language Support: Comprehend is appropriate for international applications because it supports a wide range of languages.
- Batch Processing: Batch processing can be used to analyze a lot of text data at once.
- Real-Time API: With the use of its real-time API, you can instantly evaluate text data as it is ingested into your apps.

- Integration: To create end-to-end NLP pipelines, Amazon Comprehend may be quickly linked with other AWS services like Amazon S3, AWS Glue, AWS Lambda, and more.

- HIPAA Eligibility: It can be used for healthcare-related applications that need to comply with the HIPAA, as it is HIPAA qualified.

Various applications, such as social media monitoring, customer feedback analysis, content recommendation systems, chatbots, and document categorization, frequently make use of Amazon Comprehend. It assists businesses in automating content processing, enhancing decision-making procedures based on the analysis of unstructured text, and gaining insights from textual data.

Comprehend Medical

The healthcare and life sciences sectors benefit from Amazon Comprehend Medical, a specialist NLP service provided by [19]. It is intended to extract important medical data from unstructured text, including clinical notes, physician reports, patient records, and medical literature. ML is used by Amazon Comprehend Medical to recognize and comprehend medical entities and relationships in text data.

Below are the Key features and capabilities of Amazon Comprehend Medical include:

- Entity Recognition: Comprehend Medical is able to identify and extract from text medical elements like drugs, dosage details, and treatment plans as well as medical situations (such as diseases and symptoms). This aids in automating the procedure for locating pertinent medical data.

- Attribute Detection: The service can recognize characteristics connected to medical entities, including negation (if a condition is described as ”not present”), dose frequency, and relative temporal expressions (such as ”past medical history”).

- Relationship Extraction: The relationships between the medical entities mentioned in text can be discovered using Comprehend Medical. For instance, it can show that a particular drug is recommended for a particular ailment.
- Protected Health Information (PHI) Identification: The service can identify and secure sensitive patient data, enabling healthcare businesses to comply with privacy laws like HIPAA. It can recognize and mark protected health information (PHI) within text.
- ICD-10 Code Mapping: The International Classification of Diseases, 10th Edition (ICD-10) numbers, which are frequently used for healthcare billing and coding, can be mapped to medical illnesses and clinical phrases by Comprehend Medical [9].
- RxNorm Code Mapping: Additionally, it can translate drug names into RxNorm codes, which are used in healthcare to identify drugs and their characteristics.
- Custom Entity Recognition: In order to identify specific medical entities pertinent to their use cases or domain, users might develop bespoke entity identification models.
- Batch Processing: By supporting batch processing, Comprehend Medical enables you to bulk-analyze enormous amounts of healthcare text data.
- Integration: To automate workflows and actions based on the retrieved medical data, the service can be linked with other AWS services, such as AWS Lambda.

A variety of healthcare use cases benefit from Amazon Comprehend Medical, including:

- Clinical documentation and coding automation.
- Clinical trial data extraction and analysis.
- Health insurance claims processing.

- Medical research and literature analysis.
- Disease surveillance and outbreak detection.
- Patient data extraction and population health analysis.

It makes it easier to separate structured medical data from unstructured text, saving academics and healthcare practitioners time and effort while manually reviewing and extracting data from clinical papers. It can also improve the precision and effectiveness of healthcare procedures, resulting in better patient care and research results.

Dynamo DB

A fully managed NoSQL database service, Amazon DynamoDB is offered by AWS. Its quick and predictable performance, easy scaling, and low-latency data access make it suited for a variety of applications, including online and mobile apps, gaming platforms, and IoT (Internet of Things) platforms [20].

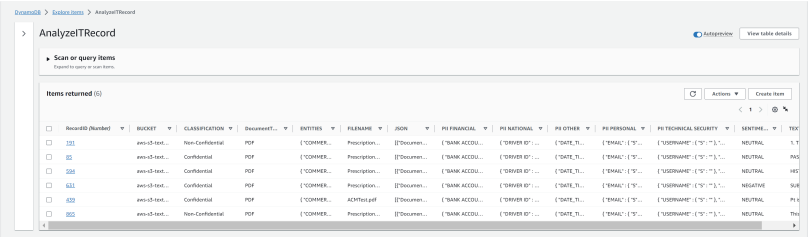

Figure 4.4: AWS DynamoDB

Below are the Key features and concepts of Amazon DynamoDB include:

- Managed Service: AWS takes care of administrative duties like hardware provisioning, setup, configuration, patching, backups, and scaling for you as DynamoDB is a fully managed database service. Developers may now concentrate on creating applications rather than running the database infrastructure thanks to this.

- NoSQL Database: Due to its NoSQL design, DynamoDB is best suited for processing unstructured, semi-structured, or structured data. It allows for flexible data modeling and supports a range of data types, including documents, key-value pairs, and JSON.

- Scalability: DynamoDB is built for smooth scaling. Without any downtime, you can start with a small capacity and increase it or decrease it as necessary. As your program expands, it may automatically divide and distribute data to guarantee dependable performance.

- Performance: The low latency and single-digit millisecond response rates offered by DynamoDB make it appropriate for use in applications that need quick data access. For storage, it makes advantage of solid-state drives (SSDs) to achieve excellent performance.

- Data Replication and High Availability: To ensure high availability and fault tolerance, DynamoDB replicates data across different Availability Zones (AZs) inside an AWS Region. Automatic repetition of this occurs.

- Data Encryption: By default, data at rest is encrypted, and you can enable it for data in transit as well. For the administration of encryption keys, DynamoDB offers AWS Key administration Service (KMS).

- Global Tables: You may replicate data across AWS Regions using DynamoDB Global Tables, providing multi-region, highly available applications with low-latency access to data in many places.

- On-Demand and Provisioned Capacity: On-demand and allocated capacity modes are options. While provisioned mode allows you to specify the required read and write capacity units, on-demand mode scales dynamically based on usage.

- Streams: DynamoDB Streams can handle events and synchronize data by capturing changes to a table's contents.
- Triggers: Triggers can be used to launch AWS Lambda functions or other AWS services automatically in response to modifications made to your DynamoDB table.
- Backup and Restore: To preserve your data, DynamoDB offers backup and restore features. Both automated and on-demand backups can be set up.
- Fine-Grained Access Control: Access to DynamoDB tables can be restricted using AWS Identity and Access Management (IAM), and fine-grained access policies can be used to enforce the restriction.

Applications including e-commerce platforms, gaming leaderboards, real-time analytics, content management systems, and IoT data storage frequently use Amazon DynamoDB. It is a strong option for developers creating applications with quickly changing data requirements due to its scalability, high availability, and low-latency access.

Amazon Comprehend service detects the PII Financial attributes like bank account number, bank routing, credit debot cvv, cre debit expiry, credit debit number, pin. It detects PII national attributes like driver id, passsport number, ssn. It detects PII Personal attributes like address, age, email, name and phone. It detects PII Technical Security like secret key, ip address, mac address, username.

Amazon Comprehend Medical service detects the Medication, Medical Condition, Protected Health Information, Test Treatment Procedure, Anatomy and Behavioral Environment Social attributes in the clinical Text and link it to medical ontologies [21].

Assumptions

- Created sample pdf and image files using the medical text samples from MTSamples.com.

- Used Amazon Textract in order to extract text from image prescriptions.
- Used Amazon Comprehend to detect PII.
- Design the DynamoDB schema to accommodate different medical entities and their relationships for auto-clasification.

Delimitations

- To Archive Healthcare Records in a Medical Cloud application using Amazon AWS Services such as S3, Textract, Comprehend, Comprehend Medical, and DynamoDB.
- To Analyze and Classify Medical Text Using AWS based on PII entities.
- To automate the extraction of medical information from documents, enabling structured storage and analysis of medical data for improved healthcare management and insights.

CHAPTER V : METHODS

Dataset

We have used the mtsamples.com Medical Transcriptions dataset [22]. The purpose of MTSamples.com is to provide you with access to a large number of transcribed medical reports. Learners and current medical transcriptions alike can use these samples to meet their daily transcription needs.

The Medical Transcriptions dataset has 140,208 sentences in the transcription column, and there are 35,805 unique words in the transcription column. There are a total of 40 categories in the original dataset, which are shown in the table 5.1.

Original Categories

MTSamples.com provides samples and reports for the following specialties. There are total 40 categories in the dataset.

Table 5.1: Original Categories.

Original Categories		
Category Number	Category Name	Number of Samples
Cat:1	Allergy / Immunology	7
Cat:2	Autopsy	8
Cat:3	Bariatrics	18

Cat:4	Cardiovascular / Pulmonary	371
Cat:5	Chiropractic	14
Cat:6	Consult - History and Phy.	516
Cat:7	Cosmetic / Plastic Surgery	27
Cat:8	Dentistry	27
Cat:9	Dermatology	29
Cat:10	Diets and Nutritions	10
Cat:11	Discharge Summary	108
Cat:12	ENT - Otolaryngology	96
Cat:13	Emergency Room Reports	75
Cat:14	Endocrinology	19
Cat:15	Gastroenterology	224
Cat:16	General Medicine	259
Cat:17	Hematology - Oncology	90
Cat:18	Hospice - Palliative Care	6
Cat:19	IME-QME-Work Comp etc.	16
Cat:20	Lab Medicine - Pathology	8

Cat:21	Letters	23
Cat:22	Nephrology	81
Cat:23	Neurology	223
Cat:24	Neurosurgery	94
Cat:25	Obstetrics / Gynecology	155
Cat:26	Office Notes	50
Cat:27	Ophthalmology	83
Cat:28	Orthopedic	355
Cat:29	Pain Management	61
Cat:30	Pediatrics - Neonatal	70
Cat:31	Physical Medicine - Rehab	21
Cat:32	Podiatry	47
Cat:33	Psychiatry / Psychology	53
Cat:34	Radiology	273
Cat:35	Rheumatology	10
Cat:36	SOAP / Chart / Progress Notes	166
Cat:37	Sleep Medicine	20

Cat:38	Speech - Language	9
Cat:39	Surgery	1088
Cat:40	Urology	156

Reduced Categories

In the original dataset categories as shown in Table 5.1, we can see that there are certain categories which have less than 50 samples highlighted as red. This is very minor as compared to the other categories which has more than 50 samples highlighted as green. We can eliminate categories with fewer than 50 samples to reduce the categories dataset as shown in Table 5.2.

Table 5.2: Reduced Categories.

Reduced Categories		
Category Number	Category Name	Number of Samples
Cat:1	Cardiovascular / Pulmonary	371
Cat:2	Consult - History and Phy.	516
Cat:3	Discharge Summary	108
Cat:4	ENT - Otolaryngology	96
Cat:5	Emergency Room Reports	75

Cat:6	Gastroenterology	224
Cat:7	General Medicine	259
Cat:8	Hematology - Oncology	90
Cat:9	Nephrology	81
Cat:10	Neurology	223
Cat:11	Neurosurgery	94
Cat:12	Obstetrics / Gynecology	155
Cat:13	Ophthalmology	83
Cat:14	Orthopedic	355
Cat:15	Pain Management	61
Cat:16	Pediatrics - Neonatal	70
Cat:17	Psychiatry / Psychology	53
Cat:18	Radiology	273
Cat:19	SOAP / Chart / Progress Notes	166
Cat:20	Surgery	1088
Cat:21	Urology	156

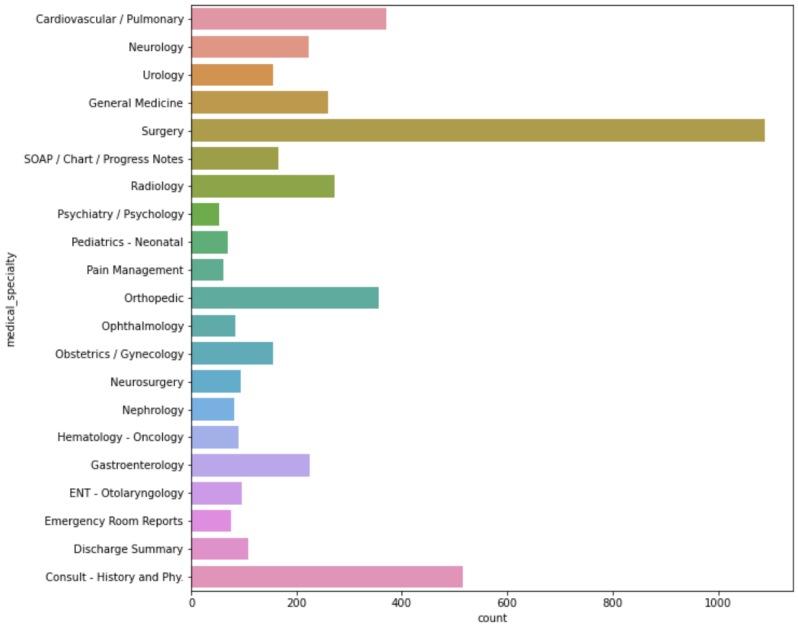

Figure 5.1: Dataset plot with reduced categories

Text Preprocessing

Text preprocessing is a method to clean the text data and make it ready to feed data to the model. Text data contains noise in various forms like emotions, punctuation, text in a different case. When we talk about human language then, there are different ways to say the same thing. In addition to removing noise from the data , the data must be turned into numerical format so the computer can process the data in an efficient way.

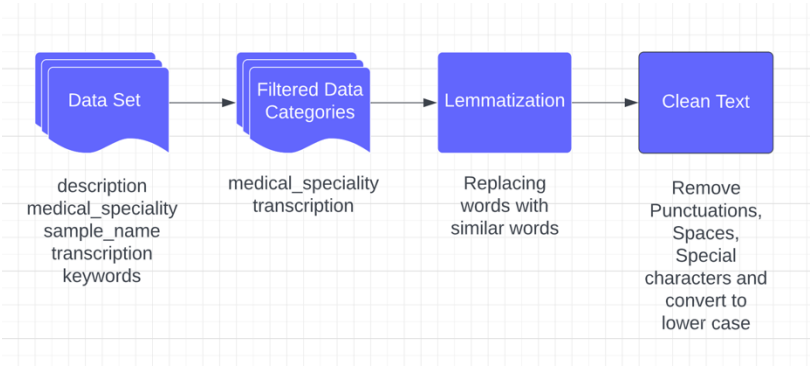

Figure 5.2: Data Pre-Processing Steps

The steps to pre-process the dataset by filtering the dataset categories, apply lemmetization and then clean the text by removing punctuations, spaces, special characters, convert to lower case etc. are shown in Figure 5.2.

Load Sample Text From Dataset

We have load sample data to see the text and perform data pre-processing. We can observe lots of noise at the first text like extra spaces, many special characters like hyphen marks and colons, different cases, and many more.

```
Sample Transcription 1: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 presentation. 3-4 weeks prior to presentation, she was found "passed out" in bed, and when awoken appeared confused, and lethargic. She
apparently recovered within 24 hours. For two weeks prior to presentation she demonstrated emotional lability, uncharacteristic of her ( outbursts of a
nger and inappropriate laughter). She left a stove on.,She began slurring her speech 2 days prior to admission. On the day of presentation she develope
d right facial weakness and began stumbling to the right. She denied any associated headache, nausea, vomiting, fever, chills, neck stiffness or visual
change. There was no history of illicit drug/ETOH use or head trauma.,PMH:, Migraine Headache.,FHX: , Unremarkable.,SHX: ,Divorced. Lives with boyfrien
d. 3 children alive and well. Denied tobacco/illicit drug use. Rarely consumes ETOH.,ROS:, Irregular menses.,EXAM: ,BP118/66.  HR83. RR 20. T36.8C.,MS:
Alert and oriented to name only. Perseverative thought processes. Utilized only one or two word answers/phrases. Non-fluent. Rarely followed commands.
Impaired writing of name.,CN: Flattened right nasolabial fold only.,Motor: Mild weakness in RUE manifested by pronator drift. Other extremities were fu
ll strength.,Sensory: withdrew to noxious stimulation in all 4 extremities.,Coordination: difficult to assess.,Station: Right pronator drift.,Gait: unr
emarkable.,Reflexes: 2/2BUE, 3/3BLE, Plantars were flexor bilaterally.,General Exam: unremarkable.,INITIAL STUDIES:, CBC, GS, UA, PT, PTT, ESR, CRP, EK
G were all unremarkable. Outside HCT showed hypodensities in the right putamen, left caudate, and at several subcortical locations (not specified).,COU
RSE: ,MRI Brian Scan, 2/11/92 revealed an old lacunar infarct in the right basal ganglia, edema within the head of the left caudate nucleus suggesting
an acute ischemic event, and arterial enhancement of the left MCA distribution suggesting slow flow. The latter suggested a vasculopathy such as Moya M
oya, or fibromuscular dysplasia. HIV, ANA, Anti-cardiolipin Antibody titer, Cardiac enzymes, TFTs, B12, and cholesterol studies were unremarkable.,She
underwent a cerebral angiogram on 2/12/92. This revealed an occlusion of the left MCA just distal to its origin. The distal distribution of the left MC
A filled on later films through collaterals from the left ACA. There was also an occlusion of the right MCA just distal to the temporal branch. Distal
branches of the right MCA filled through collaterals from the right ACA. No other vascular abnormalities were noted. These findings were felt to be aty
pical but nevertheless suspicious of a large caliber vasculitis such as Moya Moya disease. She was subsequently given this diagnosis. Neuropsychologic
testing revealed widespread cognitive dysfunction with particular impairment of language function. She had long latencies responding and understood onl
y simple questions. Affect was blunted and there was distinct lack of concern regarding her condition. She was subsequently discharged home on no medic
ations.,In 9/92 she was admitted for sudden onset right hemiparesis and mental status change. Exam revealed the hemiparesis and in addition she was fou
nd to have significant neck lymphadenopathy. OB/GYN exam including cervical biopsy, and abdominal/pelvic CT scanning revealed stage IV squamous cell ca
ncer of the cervix. She died 9/24/92 of cervical cancer.
Sample Transcription 2:ADMITTING DIAGNOSES:,1. Hematuria.,2. Benign prostatic hyperplasia.,3. Osteoarthritis.,DISCHARGE DIAGNOSES:,1. Hematuria, resolv
ed.,2. Benign prostatic hyperplasia.,3. Complex renal cyst versus renal cell carcinoma or other tumor.,4. Osteoarthritis.,HOSPITAL COURSE:, This is a 7
7-year-old African-American male who was previously well until he began having gross hematuria and clots passing through his urethra on the day of admi
ssion. He stated that he never had blood in his urine before, however, he does have a past history of BPH and he had a transurethral resection of prost
ate more than 10 years ago. He was admitted to a regular bed. Dr. G of Urology was consulted for evaluation of his hematuria. During the workup for thi
s, he had a CT of the abdomen and pelvis with and without contrast with early and late-phase imaging for evaluation of the kidneys and collecting syste
m. At that time, he was shown to have multiple bilateral renal cysts with one that did not meet classification as a simple cyst and ultrasound was reco
mmended.,He had an ultrasound done of the cyst which showed a 2.1 x 2.7 cm mass arising from the right kidney which, again, did not fit ultrasound crit
eria for a simple cyst and they recommended further evaluation by an MRI as this could be a hemorrhagic cyst or a solid mass or tumor, so an MRI was sc
heduled on the day of discharge for further evaluation of this. The report was not back at discharge. The patient had a cystoscopy and transurethral re
section of prostate as well with entire resection of the prostate gland. Pathology on this specimen showed multiple portions of prostatic tissue which
was primarily fibromuscular, and he was diagnosed with nonprostatic hyperplasia. His urine slowly cleared. He tolerated a regular diet with no difficul
ties in his activities of daily living, and his Foley was removed on the day of discharge.,He was started on ciprofloxacin, Colace, and Lasix after the
transurethral resection and continued these for a short course. He is asked to continue the Colace as an outpatient for stool softening for comfort.,DI
SCHARGE MEDICATIONS:, Colace 100 mg 1 b.i.d.,DISCHARGE FOLLOWUP PLANNING:, The patient is to follow up with his primary care physician at ABCD, Dr. B o
r Dr. J, the patient is unsure of which, in the next couple weeks. He is to follow up with Dr. G of  Urology in the next week by phone in regards to th
e patient's MRI and plans for a laparoscopic partial renal resection biopsy. This is scheduled for the week after discharge potentially by Dr. G, and t
he patient will discuss the exact time later this week. The patient is to return to the emergency room or to our clinic if he has worsening hematuria a
gain or no urine output.
Sample Transcription 3:PREOPERATIVE DIAGNOSES: , Phimosis and adhesions.,POSTOPERATIVE DIAGNOSES:  ,Phimosis and adhesions.,PROCEDURES PERFORMED: , Cir
cumcision and release of ventral chordee.,ANESTHESIA:  ,Local MAC.,ESTIMATED BLOOD LOSS: , Minimal.,FLUIDS: , Crystalloid.  The patient was given antib
iotics preop.,BRIEF HISTORY: , This is a 43-year-old male who presented to us with significant phimosis, difficulty retracting the foreskin.  The patie
nt had buried penis with significant obesity issues in the suprapubic area.  Options such as watchful waiting, continuation of slowly retracting the sk
in, applying betamethasone cream, and circumcision were discussed.  Risk of anesthesia, bleeding, infection, pain, MI, DVT, PE, and CVA risks were disc
ussed.  The patient had discussed this issue with Dr Khan and had been approved to get off of the Plavix.  Consent had been obtained.  Risk of scarrin
g, decrease in penile sensation, and unexpected complications were discussed.  The patient was told about removing the dressing tomorrow morning, okay
to shower after 48 hours, etc.  Consent was obtained.,DESCRIPTION OF PROCEDURE:  ,The patient was brought to the OR.  Anesthesia was applied.  The pati
ent was placed in supine position.  The patient was prepped and draped in usual sterile fashion.  Local MAC anesthesia was applied.  After draping, 17
mL of mixture of 0.25% Marcaine and 1% lidocaine plain were applied around the dorsal aspect of the penis for dorsal block.  The patient had significan
t phimosis and slight ventral chordee.  Using marking pen, the excess foreskin was marked off.  Using a knife, the ventral chordee was released.  The u
rethra was intact.  The excess foreskin was removed.  Hemostasis was obtained using electrocautery.  A 5-0 Monocryl stitches were used for 4 interrupte
d stitches and horizontal mattresses were done.  The patient tolerated the procedure well.  There was excellent hemostasis.  The penis was straight.  V
aseline gauze and Kerlix were applied.  The patient was brought to the recovery in stable condition.  Plan was for removal of the dressing tomorrow.  O
kay to shower after 48 hours.
```

Figure 5.3: Sample Transcriptions.

Remove Special Characters and Symbols

We have used python regular expression library to find the special characters and replace them by space. The function used to compile the regular expression patters in re.compile(pattern, flags=0) to match the special characters and then use the function re.sub(pattern,

replacement, string, count=0, flags=0) to return the string obtained by replacing the left-most non-overlapping occurrences of pattern in string by the replacement by a blank space.

Lemmatization

Lemmatization is used to stem the words into root word but differs in working. Actually, Lemmatization is a systematic way to reduce the words into their lemma by matching them with a language dictionary. We have used Wordnet Lemmatizer with NLTK. Wordnet is an large, freely and publicly available lexical database for the English language aiming to establish structured semantic relationships between words. It offers lemmatization capabilities as well and is one of the earliest and most commonly used lemmatizers. NLTK offers an interface to it, but you have to download it first in order to use it. Follow the below instructions to install nltk and download wordnet. In order to lemmatize, you need to create an instance of the WordNetLemmatizer() and call the lemmatize() function on a single word.

Clean Text

We have converted the text to lowercase. If the text is in the same case, it is easy for a machine to interpret the words because the lower case and upper case are treated differently by the machine. One of the other text processing techniques is removing punctuation's. We can directly remove punctuation from string by using the translate method. The translate method produces a string in which some characters are substituted with characters from a dictionary or a mapping table. The join method is also used to remove the punctuations. The join method allows us to create strings from iterable objects in various ways. It concatenates each component of an iterable.

```
def clean_text(text ):
    text = text.translate(str.maketrans('', '', string.punctuation))
    text1 = ''.join([w for w in text if not w.isdigit()])
    RE_REPLACE_BY_SPACE = re.compile('[/(){}\[\]\|@,;]')

    text2 = text1.lower()
    text2 = RE_REPLACE_BY_SPACE.sub('', text2)
    return text2

def lemmatize_text(text):
    wordlist=[]
    lemmatizer = WordNetLemmatizer()
    sentences=sent_tokenize(text)

    intial_sentences= sentences[0:1]
    final_sentences = sentences[len(sentences)-2: len(sentences)-1]

    for sentence in intial_sentences:
        words=word_tokenize(sentence)
        for word in words:
            wordlist.append(lemmatizer.lemmatize(word))
    for sentence in final_sentences:
        words=word_tokenize(sentence)
        for word in words:
            wordlist.append(lemmatizer.lemmatize(word))
    return ' '.join(wordlist)
```

Figure 5.4: Code Snippet for Clean Text and Lemmetization.

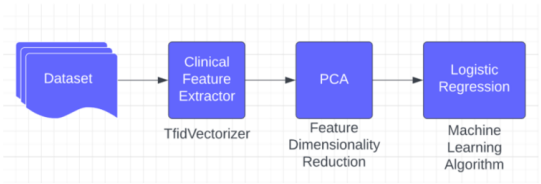

Figure 5.5: First Method: Using TFidVectorizer, PCA and Logistic Regression Machine Learning Model

First Method: TFidVectorizer, PCA and Logistic Regression

TF-IDF Vectorizer

TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. Let's take an example, we have a string or Bag of Words (BOW) and we have to extract information from it, then we can use this approach.

```
vectorizer = TfidfVectorizer(analyzer='word', stop_words='english',
    ngram_range=(1,3), max_df=0.75, use_idf=True, smooth_idf=True,
    max_features=1000)
tfIdfMat  = vectorizer.fit_transform(data['transcription'].tolist() )
feature_names = sorted(vectorizer.get_feature_names())
print(feature_names)
```

Figure 5.6: Code Snippet for TF-IDF Vectorizer.

The tf-idf value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and The inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents. TF-IDF can be computed as tf * idf.

['abc', 'abcd', 'abdomen', 'abdomen pelvis', 'abdominal', 'abdominal pain', 'abnormal', 'abscess', 'activity', 'acute', 'additional', 'adenocarcinoma',
'adequate', 'administered', 'admission', 'admitted', 'admitting', 'africanamerican', 'age', 'ago', 'airway', 'alcohol', 'anemia', 'anesthesia', 'anesth
esia care', 'anesthesia general', 'anesthesia general endotracheal', 'anesthesia local', 'anesthesia wa', 'anesthetic', 'angina', 'angiography', 'ankl
e', 'anterior', 'antibiotic', 'aortic', 'apnea', 'apparent', 'appendicitis', 'applied', 'appointment', 'appropriate', 'approximately', 'area', 'arm',
'artery', 'artery disease', 'arthritis', 'arthroplasty', 'asked', 'aspect', 'aspiration', 'assessment', 'associated', 'atrial', 'atrial fibrillation',
'awakened', 'axial', 'axial ct', 'axial ct image', 'axis', 'baby', 'base', 'benefit', 'benign', 'better', 'bid', 'bilateral', 'bilaterally', 'biopsy',
'bladder', 'bleeding', 'block', 'blood', 'blood loss', 'blood loss cc', 'blood loss minimal', 'blood loss ml', 'blood loss wa', 'blood pressure', 'bod
y', 'bone', 'bowel', 'boy', 'brain', 'breast', 'breath', 'breathing', 'brief', 'brief history', 'brief history patient', 'bronchoscopy', 'brought', 'br
ought operating', 'brought operating room', 'bunion', 'bypass', 'cancer', 'carcinoma', 'cardiac', 'care', 'care unit', 'carotid', 'carpal', 'carpal tun
nel', 'carpal tunnel syndrome', 'case', 'cataract', 'cataract right', 'cataract right eye', 'catheter', 'catheterization', 'caucasian', 'caucasian fema
le', 'caucasian male', 'cc', 'cc cc', 'cell', 'cell carcinoma', 'central', 'cervical', 'cervical spine', 'chamber', 'change', 'check', 'chest', 'chest
pain', 'chief', 'chief complaint', 'child', 'cholecystectomy', 'cholecystitis', 'cholelithiasis', 'chronic', 'circumcision', 'clear', 'clinic', 'clinic
al', 'closed', 'closed vicryl', 'closure', 'cm', 'colon', 'colonoscopy', 'come', 'common', 'complaining', 'complaint', 'complete', 'completed', 'comple
x', 'complication', 'complications', 'complications estimated', 'complications estimated blood', 'compression', 'concern', 'condition', 'consent', 'con
sent wa', 'consent wa obtained', 'consistent', 'consult', 'consultation', 'continue', 'continued', 'contrast', 'contrast reason', 'contrast reason exa
m', 'control', 'cord', 'coronal', 'coronary', 'coronary artery', 'coronary artery disease', 'correct', 'cough', 'count', 'count correct', 'course', 'cr
ystalloid', 'ct', 'ct abdomen', 'ct image', 'ct scan', 'current', 'cyst', 'cystoscopy', 'daily', 'data', 'date', 'day', 'days', 'deep', 'defect', 'defo
rmity', 'degenerative', 'delivery', 'denies', 'dental', 'department', 'depression', 'descending', 'description', 'description procedure', 'description
procedure patient', 'developed', 'diabetes', 'diabetes mellitus', 'diagnosed', 'diagnoses', 'diagnoses patient', 'diagnoses patient wa', 'diagnoses spo
nge', 'diagnoses wa', 'diagnosis', 'diagnosis acute', 'diagnosis bilateral', 'diagnosis cataract', 'diagnosis cervical', 'diagnosis chronic', 'diagnosi
s es', 'diagnosis left', 'diagnosis recurrent', 'diagnosis right', 'diagnostic', 'diameter', 'diarrhea', 'did', 'difficulty', 'direct', 'disc', 'discec
tomy', 'discharge', 'discharge diagnoses', 'discharge diagnosis', 'discharged', 'discomfort', 'discussed', 'discussed patient', 'disease', 'disk', 'dis
order', 'disposition', 'disposition patient', 'distal', 'distress', 'doe', 'doing', 'dorsal', 'dr', 'drain', 'drainage', 'drains', 'draped', 'draped st
erile', 'draped usual', 'dressing', 'dressing applied', 'dressing wa', 'dressing wa applied', 'drop', 'dysphagia', 'ear', 'echocardiogram', 'edema', 'e
ffusion', 'ekg', 'elbow', 'elevated', 'emergency', 'emergency department', 'emergency room', 'end', 'end procedure', 'endoscopy', 'endotracheal', 'endo
tracheal anesthesia', 'endotracheal tube', 'endstage', 'endstage renal', 'endstage renal disease', 'epidural', 'episode', 'es', 'esophageal', 'estimate
d', 'estimated blood', 'estimated blood loss', 'evaluate', 'evaluation', 'evidence', 'exam', 'exam ct', 'exam mri', 'examination', 'excision', 'exercis
e', 'explained', 'explained patient', 'external', 'extremity', 'extubated', 'eye', 'eye postoperative', 'eye postoperative diagnosis', 'eye procedure',
'face', 'failure', 'fall', 'family', 'family history', 'fascia', 'fashion', 'felt', 'female', 'female history', 'female present', 'female wa', 'femora
l', 'fetal', 'fever', 'fibrillation', 'final', 'finding', 'findings', 'findings patient', 'finger', 'fixation', 'flow', 'fluid', 'fluids', 'foley', 'fo
llow', 'followed', 'following', 'followup', 'foot', 'foramen', 'foreign', 'foreign body', 'fracture', 'free', 'french', 'frontal', 'function', 'fusio
n', 'general', 'general anesthesia', 'general endotracheal', 'general endotracheal anesthesia', 'gentleman', 'gi', 'given', 'going', 'good', 'good cond
ition', 'grade', 'graft', 'greater', 'gross', 'ha', 'ha history', 'hand', 'hardware', 'having', 'head', 'headache', 'health', 'hearing', 'heart', 'hear
t rate', 'help', 'hematoma', 'hemorrhage', 'hemostasis', 'hernia', 'herniated', 'herniated nucleus', 'herniated nucleus pulposus', 'high', 'hip', 'hist
t rate', 'help', 'hematoma', 'hemorrhage', 'hemostasis', 'hernia', 'herniated', 'herniated nucleus', 'herniated nucleus pulposus', 'high', 'hip', 'hist
ory', 'history patient', 'history patient yearold', 'history present', 'history present illness', 'history yearold', 'home', 'hospital', 'hour', 'hx',
'hx yo', 'hx yo rhf', 'hx yo rhm', 'hypertension', 'hypertrophy', 'identified', 'ii', 'illness', 'illness patient', 'illness patient pleasant', 'illnes
s patient yearold', 'illness yearold', 'illness yearold female', 'illness yearold male', 'illness yearold white', 'image', 'imaging', 'implantation',
'impression', 'incision', 'incision wa', 'including', 'increased', 'indication', 'indication surgery', 'indications', 'indications patient', 'indicatio
ns patient yearold', 'indications procedure', 'indications procedure patient', 'indications surgery', 'infection', 'inferior', 'informed', 'informed co
nsent', 'informed consent wa', 'inguinal', 'inguinal hernia', 'initially', 'injected', 'injection', 'injury', 'insertion', 'instructed', 'instruction',
'instrument', 'intact', 'internal', 'interpretation', 'interrupted', 'intervention', 'intraocular', 'intraocular lens', 'intraoperative', 'intravenou
s', 'irrigated', 'iv', 'joint', 'kidney', 'knee', 'known', 'laceration', 'lady', 'lap', 'laparoscopic', 'large', 'lateral', 'layer', 'left', 'left brea
st', 'left foot', 'left hip', 'left knee', 'left lower', 'left lower extremity', 'left shoulder', 'left upper', 'leg', 'length', 'lens', 'lesion', 'lev
el', 'lidocaine', 'ligament', 'likely', 'liver', 'll', 'lobe', 'local', 'loss', 'loss cc', 'loss minimal', 'loss ml', 'loss wa', 'low', 'low pain', 'lo
wer', 'lower extremity', 'lower quadrant', 'ls', 'lumbar', 'lung', 'lymph', 'lymph node', 'mac', 'male', 'male present', 'male wa', 'man', 'managemen
t', 'manner', 'marcaine', 'mass', 'medial', 'medical', 'medical history', 'medication', 'medications', 'medium', 'mellitus', 'membrane', 'mental', 'met
astatic', 'mg', 'middle', 'mild', 'minimal', 'minute', 'minutes', 'ml', 'mm', 'moderate', 'mom', 'monitored', 'monocryl', 'month', 'monthold', 'mornin
g', 'mother', 'motor', 'mouth', 'mr', 'mri', 'ms', 'multiple', 'muscle', 'myocardial', 'nasal', 'nausea', 'nausea vomiting', 'neck', 'need', 'needed',
'needle', 'needle count', 'needle count correct', 'negative', 'nerve', 'neurologic', 'new', 'night', 'node', 'noncontrast', 'normal', 'nose', 'note',
'noted', 'nuclear', 'nucleus', 'nucleus pulposus', 'numbness', 'obstruction', 'obstructive', 'obtained', 'obtained patient', 'obtained patient wa', 'ob
taining', 'office', 'old', 'onset', 'open', 'operating', 'operating room', 'operating room placed', 'operating table', 'operation', 'operation patien
t', 'operation performed', 'operations', 'operative', 'operative procedure', 'oral', 'osteoarthritis', 'otitis', 'outpatient', 'ox', 'pacemaker', 'pac
u', 'pain', 'pain history', 'pain history present', 'parent', 'partial', 'past', 'past medical', 'past medical history', 'patent', 'pathology', 'patien
t', 'patient ha', 'patient monthold', 'patient pleasant', 'patient pleasant yearold', 'patient present', 'patient tolerated', 'patient tolerated proced
ure', 'patient wa', 'patient wa brought', 'patient wa extubated', 'patient wa placed', 'patient wa taken', 'patient yearold', 'patient yearold africana
merican', 'patient yearold caucasian', 'patient yearold female', 'patient yearold gentleman', 'patient yearold male', 'patient yearold white', 'patient
yearold woman', 'pelvic', 'pelvis', 'percutaneous', 'performed', 'performed patient', 'perfusion', 'period', 'persistent', 'phacoemulsification', 'phys
ical', 'physical examination', 'physician', 'place', 'placed', 'placed supine', 'placed supine position', 'placement', 'plan', 'plate', 'pleasant', 'pl
easant yearold', 'pleural', 'pleural effusion', 'pneumonia', 'point', 'polyp', 'position', 'positive', 'possible', 'post', 'posterior', 'postoperativ
e', 'postoperative diagnoses', 'postoperative diagnosis', 'postoperative diagnosis left', 'postoperative diagnosis right', 'postprocedure', 'pregnanc
y', 'preoperative', 'preoperative diagnoses', 'preoperative diagnoses patient', 'preoperative diagnoses sponge', 'preoperative diagnosis', 'preoperativ
e diagnosis acute', 'preoperative diagnosis bilateral', 'preoperative diagnosis left', 'preoperative diagnosis right', 'prepped', 'prepped draped', 'pr
epped draped usual', 'preprocedure', 'present', 'present illness', 'present illness patient', 'present illness yearold', 'present today', 'presentatio
n', 'presented', 'presented emergency', 'presented emergency room', 'pressure', 'previous', 'previously', 'primary', 'prior', 'prn', 'problem', 'proced
ure', 'procedure informed', 'procedure informed consent', 'procedure left', 'procedure patient', 'procedure patient wa', 'procedure patient yearold',
'procedure performed', 'procedure right', 'procedure wa', 'procedure yearold', 'procedures', 'procedures performed', 'progressive', 'prostate', 'prosta
te cancer', 'protocol', 'proximal', 'pulmonary', 'pulposus', 'quadrant', 'question', 'radial', 'radiation', 'radiculopathy', 'rate', 'reason', 'reason
consult', 'reason consultation', 'reason exam', 'reason visit', 'received', 'recent', 'recently', 'recommendation', 'recommended', 'reconstruction', 'r
ecovery', 'recovery room', 'recovery room satisfactory', 'recovery room stable', 'recovery stable', 'rectal', 'recurrent', 'reduction', 'referral', 're
ferred', 'reflux', 'regarding', 'region', 'related', 'release', 'removal', 'removed', 'renal', 'renal disease', 'renal mass', 'repair', 'replacement',
'report', 'resection', 'residual', 'respiratory', 'result', 'return', 'returned', 'revealed', 'review', 'reviewed', 'rhf', 'rhm', 'right', 'right breas
t', 'right eye', 'right eye postoperative', 'right foot', 'right inguinal', 'right inguinal hernia', 'right knee', 'right lower', 'right shoulder', 'ri
ght upper', 'righthanded', 'risk', 'risk benefit', 'room', 'room placed', 'room placed supine', 'room satisfactory', 'room satisfactory condition', 'ro
om stable', 'room stable condition', 'routine', 'rule', 'running', 'ruptured', 'satisfactory', 'satisfactory condition', 'scan', 'scope', 'screening',
'screw', 'second', 'secondary', 'sedation', 'seen', 'seizure', 'sent', 'service', 'severe', 'shortness', 'shortness breath', 'shoulder', 'showed', 'shu
nt', 'sign', 'significant', 'signs', 'single', 'sinus', 'site', 'size', 'skin', 'skin wa', 'sleep', 'small', 'soft', 'soft tissue', 'space', 'specime
n', 'specimens', 'speech', 'spinal', 'spine', 'spondylosis', 'sponge', 'sponge lap', 'sponge needle', 'sponge needle count', 'squamous', 'squamous cel
l', 'squamous cell carcinoma', 'stable', 'stable condition', 'stage', 'standard', 'started', 'state', 'status', 'status post', 'stenosis', 'stent', 'st
erile', 'sterile dressing', 'sterile dressing applied', 'sterile fashion', 'steristrips', 'stomach', 'stone', 'stress', 'stress test', 'stroke', 'stud
y', 'subcutaneous', 'subcutaneous tissue', 'subcuticular', 'subdural', 'subdural hematoma', 'subglottic', 'subjective', 'subjective patient', 'subjecti
ve patient yearold', 'subjective yearold', 'suite', 'summary', 'superior', 'supine', 'supine operating', 'supine position', 'surgery', 'surgery patien
t', 'surgical', 'suture', 'swelling', 'symptom', 'symptomatic', 'syndrome', 'table', 'taken', 'taken operating', 'taken operating room', 'taken recover
y', 'taken recovery room', 'tear', 'technique', 'temporal', 'tendon', 'test', 'testis', 'therapy', 'thoracic', 'thyroid', 'tibial', 'time', 'tissue',
'title', 'title operation', 'today', 'tolerated', 'tolerated procedure', 'tolerated procedure wa', 'topical', 'total', 'total knee', 'tourniquet', 'tou
rniquet time', 'transferred', 'transferred recovery', 'treated', 'treatment', 'tube', 'tumor', 'tunnel', 'tunnel syndrome', 'type', 'ultrasound', 'unde
rwent', 'unit', 'unremarkable', 'upper', 'upper extremity', 'upper lobe', 'urinary', 'urine', 'use', 'used', 'using', 'usual', 'usual fashion', 'valv
e', 'vein', 'venous', 'ventricular', 'versed', 'vicryl', 'vicryl suture', 'view', 'vision', 'visit', 'visual', 'vital', 'vital signs', 'vomiting', 'w
a', 'wa admitted', 'wa applied', 'wa awakened', 'wa brought', 'wa brought operating', 'wa closed', 'wa extubated', 'wa given', 'wa injected', 'wa note
d', 'wa obtained', 'wa obtained patient', 'wa performed', 'wa placed', 'wa prepped', 'wa prepped draped', 'wa referred', 'wa removed', 'wa seen', 'wa t
aken', 'wa taken operating', 'wa taken recovery', 'wa transferred', 'wa used', 'wall', 'weakness', 'week', 'weight', 'white', 'white female', 'white ma
le', 'woman', 'work', 'worsening', 'wound', 'wound wa', 'wrist', 'xray', 'xyz', 'year', 'year ago', 'year old', 'yearold', 'yearold africanamerican',
'yearold boy', 'yearold caucasian', 'yearold female', 'yearold female present', 'yearold gentleman', 'yearold male', 'yearold man', 'yearold white', 'y
earold white female', 'yearold white male', 'yearold woman', 'yo', 'yo rhf', 'yo rhm']

Figure 5.7: Sample Features using TF-IDF Vectorizer

Transformer vs. Tfidfvectorizer

The goal of Tfidftransformer and Tfidfvectorizer from Scikit-learn is the same: to transform a set of raw documents into a matrix of TF-IDF features.

IDF-TFThe Vectorizer calculates a word's significance within a document in relation to a group of documents. This method is useful for figuring out a term's importance within

a corpus of documents. TF-IDF generates sparse feature vectors, each of which indicates how important a phrase is in a given document. These vectors fit with conventional statistical models and can be easily understood.It is excellent for situations where feature interpretability and computational speed are important because it is simple and computationally efficient.

Transformer-based models belong to the deep learning domain and are specifically made for language production and understanding problems. Examples of these models are BERT, GPT, and their variations. By utilizing attention mechanisms, these models are able to comprehend the context and semantic connections among words in a text, effectively capturing intricate patterns and subtleties. Large text corpora are frequently used to pre-train them, which makes transfer learning possible for jobs in a variety of industries, including healthcare. The performance of these models can be greatly enhanced by fine-tuning them with domain-specific data. Transformer models may need more data to be trained because they are more computationally demanding. Deep learning models' complicated architecture and ”black-box” nature may make them difficult to interpret.

TF-IDFVectorizer is recommended for healthcare data where feature interpretability is critical. Transformer-based models may perform better when dealing with complex text that has context-dependent semantics or when a deeper comprehension of the links between medical terminology is needed. For transformer models to train efficiently, substantial computer resources and greater datasets are needed. Given its lower resource requirements, TF-IDFVectorizer may be a better fit for smaller datasets.

A hybrid strategy can also work well in many situations. Leverage the interpretability of TF-IDFVectorizer and the contextual knowledge of transformers by fine-tuning them for certain healthcare tasks and exploiting their pre-trained models. The decision ultimately comes down to the particulars of the medical data and the analysis's goals.

we have used IDFVectorizer for this research.

PCA

As you can see in Figure 5.7, we are dealing with about lots of different features. These are all the distinct words found in our corpus. We can transform and reduce our input feature space through the Principal Component Analysis (PCA).

```
pca = PCA(n_components=0.95)
tfIdfMat_reduced = pca.fit_transform(tfIdfMat.toarray())
labels = data['medical_specialty'].tolist()
category_list = data.medical_specialty.unique()
```

Figure 5.8: PCA Decompostion Technique.

Principal component analysis, or PCA, is a popular method for analyzing large datasets with a lot of dimensions or features per observation, making the data easier to understand while keeping as much information as possible and showing multidimensional data. PCA is formally a statistical method for reducing a dataset's dimensionality. By linearly transforming the data into a new coordinate system with fewer dimensions than the initial data, the majority of the variation in the data can be described. In many studies, the first two principal components are used to visually identify clusters of closely related data points and plot the data in two dimensions.

t-SNE Plot

t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. It is a nonlinear dimensionality reduction technique for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by dis-

tant points with high probability. As you can see in Figure 5.9, there is lot of overlapping among the features.

Figure 5.9: Initial TSNE Plot

Logistic Regression Classifier

We have used the Logistic Regression Classifier with the below hyper-parameters as shown in Figure 3.10. The logistic regression classifier uses the weighted combination of the input features and passes them through a sigmoid function.

```
X_train, X_test, y_train, y_test = train_test_split(tfIdfMat_reduced,
    labels, stratify=labels,random_state=1)
print('Train_Set_Size:'+str(X_train.shape))
print('Test_Set_Size:'+str(X_test.shape))

clf = LogisticRegression(penalty= 'elasticnet', solver= 'saga', l1_ratio
    =0.5, random_state=1).fit(X_train, y_train)
y_test_pred= clf.predict(X_test)
```

Figure 5.10: Logistic Regression Model.

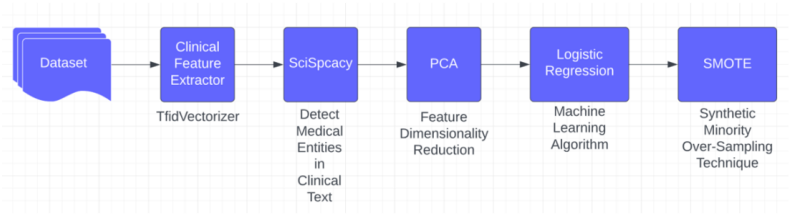

Figure 5.11: Second Method Using TFidVectorizer, SciSpacy, PCA, Logistic Regression and SMOTE.

Second Method: SciSpacy, Logistic Regression and SMOTE

The MT samples dataset is highly imbalanced as surgery category has significantly higher samples (i.e., 1088) as compared to other categories. The imbalanced dataset is the problem where data belonging to one class is significantly higher or lower than that belonging to other classes. The simplest way to fix imbalanced dataset is simply balancing them by oversampling instances of the minority class or undersampling instances of the majority class. Using advanced techniques like SMOTE helps to create new synthetic instances from minority class.

We have used second approach by using Scispacy, Logistic Regression and SMOTE as shown in Figure 5.11 in order to resolve data imbalance problem by applying domain knowledge like surgery category is superset of other categories. SOAP / Chart / Progress Notes, Office Notes, Consult - History and Phy., Emergency Room Reports and Discharge Summary are superset of certain specialities classes like Cardiology, Gastroenterology and Gynecology. General Medicine and Pain Management is also superset of our data. We can remove these categories. Also, Neurosurgery is related to Neurology and Nephrology is related to Urology. We can combine related categories.

SciSpacy

A full spaCy pipeline and models for scientific/biomedical documents using NLP. SciSpacy is a powerful tool, especially NER, or identifying keywords (known as entities) and

ordering them into categories. We have used scispacy model in order to process the medical text and get the relevent features.

SMOTE

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class [10]. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together. Since some classes are in minority, we can use SMOTE to generate more sample form minority class to solve the data imbalance problem.

```
smote_over_sample = SMOTE(sampling_strategy='minority')
labels = data['medical_specialty'].tolist()
X, y = smote_over_sample.fit_resample(tfIdfMat_reduced, labels)

X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y,
    random_state=1)
print('Train_Set_Size:'+str(X_train.shape))
print('Test_Set_Size:'+str(X_test.shape))

clf = LogisticRegression(penalty= 'elasticnet', solver= 'saga', l1_ratio
    =0.5, random_state=1).fit(X_train, y_train)
y_test_pred= clf.predict(X_test)
```

Figure 5.12: Model construct of a Logisic Regression classifier using SMOTE.

CHAPTER VI : RESULTS

Healthcare Text Analytics Using ML Techniques

Here, We have compared results from the two approaches explained in the methods sections. First approach is using the logistic regression to handle multi-class classification problems in healthcare and second approach is applying domain knowledge with SMOTE technique. We have evaluated the models using appropriate metrics like accuracy, precision, recall, and F1-score for multi-class classification.

Initial Results

Our initial results are not good as you can see that diagonal is not aligned properly in the confusion matrix as shown in Figure 6.1. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The lower the diagonal values of the confusion matrix, indicating many incorrect predictions. The initial results with reduced categories expressed in precision, recall, f1-score, and support metrics as shown in Figure 6.2. The overall accuracy is 38 percent only which explains that the model has difficulties in distinguishing the medical documents.

Final Results

Our final results are improved now. The diagonal elements in the confusion matrix which represents the number of points for which the predicted label is equal to the true label is higher now, indicating many correct predictions as shown in Figure 6.3. Final Results expressed in precision, recall, f1-score and support metrics using SMOTE as shown in Figure 6.6. Overall accuracy is improved to 70 percent now. Precision is highest for ophthalmology and Recall is highest for Psychology.

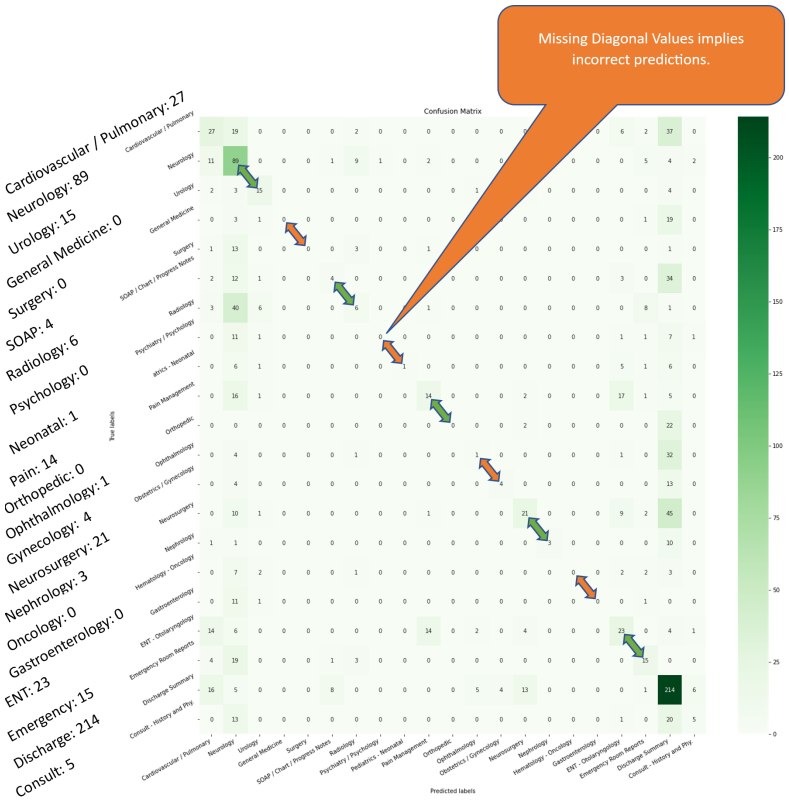

Figure 6.1: Initial Confusion Matrix for the DataSet

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.33	0.29	0.31	93
Neurology	0.42	0.25	0.31	56
Urology	0.33	0.13	0.19	39
General Medicine	0.24	0.09	0.13	65
Surgery	0.44	0.79	0.57	272
SOAP / Chart / Progress Notes	0.38	0.36	0.37	42
Radiology	0.34	0.34	0.34	68
Psychiatry / Psychology	0.00	0.00	0.00	13
Pediatrics - Neonatal	0.00	0.00	0.00	17
Pain Management	1.00	0.20	0.33	15
Orthopedic	0.43	0.24	0.30	89
Ophthalmology	0.50	0.19	0.28	21
Obstetrics / Gynecology	0.11	0.03	0.04	39
Neurosurgery	0.00	0.00	0.00	24
Nephrology	1.00	0.05	0.10	20
Hematology - Oncology	0.00	0.00	0.00	22
Gastroenterology	0.29	0.07	0.11	56
ENT - Otolaryngology	0.00	0.00	0.00	24
Emergency Room Reports	0.00	0.00	0.00	19
Discharge Summary	0.50	0.56	0.53	27
Consult - History and Phy.	0.30	0.69	0.42	129
accuracy			0.38	1150
macro avg	0.32	0.20	0.21	1150
weighted avg	0.35	0.38	0.32	1150

Figure 6.2: Initial Results for the DataSet

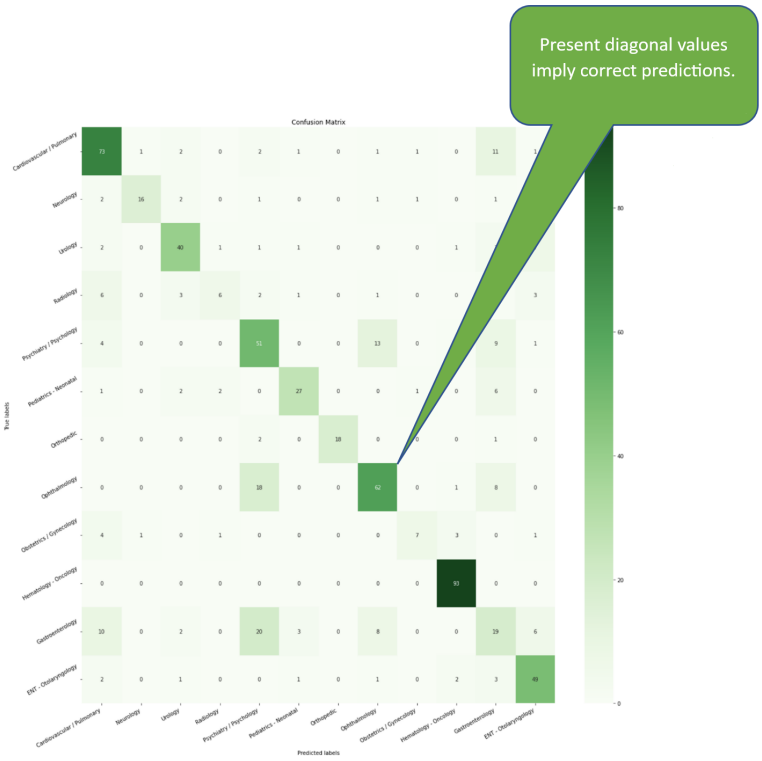

Figure 6.3: Final Confusion Matrix using SMOTE

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.70	0.78	0.74	93
Neurology	0.53	0.65	0.58	79
Urology	0.72	0.83	0.77	59
Radiology	0.31	0.28	0.29	68
Psychiatry / Psychology	0.92	1.00	0.96	93
Pediatrics - Neonatal	0.70	0.41	0.52	17
Orthopedic	0.71	0.70	0.70	89
Ophthalmology	1.00	0.86	0.92	21
Obstetrics / Gynecology	0.79	0.69	0.74	39
Hematology - Oncology	0.60	0.26	0.36	23
Gastroenterology	0.77	0.71	0.74	56
ENT - Otolaryngology	0.89	0.67	0.76	24
accuracy			0.70	661
macro avg	0.72	0.65	0.67	661
weighted avg	0.70	0.70	0.69	661

Figure 6.4: Final Results using SMOTE

Discussion

- Logistic Regression a suitable choice for various multi-class classification scenarios related to medical texts. It is simple to implement and easy to comprehend.
- Handle high-dimensional data, such as text data with a large number of different words or attributes.
- Faster and more accurate as compared to human classification.
- If we have imbalanced dataset, we can use SMOTE Algorithm to create synthetic samples for minority classes to improve the results.
- We can do custom feature extraction to the results and the major thing is here that domain knowledge plays an important role in such classification problems.

Data Classification Using AWS Cloud ML Services

The Medical text can be classified into two categories i.e., confidential and non-confidential based on the PHI information present in the medical text. We have uploaded few sample prescriptions through our Medical ArchiveIT and AnalyzeIT App into the S3 bucket and made a call to AWS Textract service to extract the text, then made a call to AWS Comprehend and ACM service in order to detect PII and PHI information. In case it contains these entities, we have written a custom rule in python to classify it as a confidential or a non-confidential document and store this information in DynamoDB database.

Results

Here, we can see that the extracted text from Prescription1.pdf is successfully classified into Confidential document as it contains PHI information.

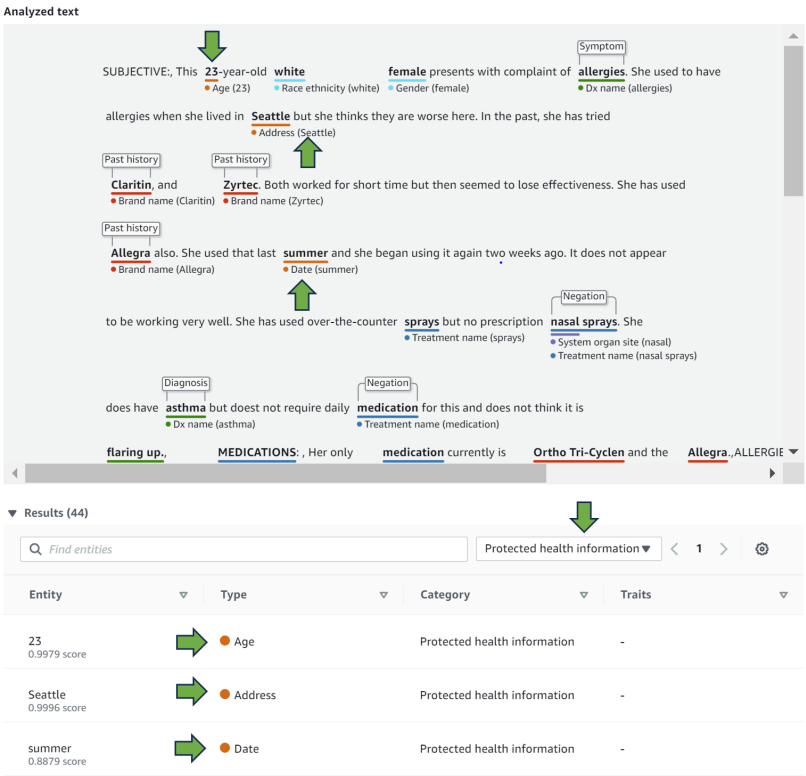

Figure 6.5: PHI Results using AWS Comprehend Medical

Record ID	File Name	Classification
191	Prescription5.pdf	Non-Confidential
85	Prescription2.pdf	Confidential
594	Prescription3.pdf	Confidential
631	Prescription1.pdf	Confidential
439	ACMTest.pdf	Confidential
865	Prescription6.pdf	Non-Confidential

Figure 6.6: Data Classification using AWS ML Services

Discussion

- This solution is good for innovation in healthcare, offering opportunities for developing new archival applications, improving patient care, and advancing medical research for analysis.

- Comprehend Medical might extract medical entities from the text but might lack a comprehensive understanding of context, such as negations or complex medical relationships, which might require additional processing for accurate analysis, espe-

cially when dealing with critical medical data or in contexts where high accuracy is required.

- Here, we classified the records into 2 categories, in case, there are more categories, we might need to write custom rules based on the confidence score of entities extraction to classify the records.

CHAPTER VII : CONCLUSIONS AND FUTURE WORK

Conclusions

This medical transcriptions data set is quite noisy and the text in the transcriptions overlaps with the categories. Figure 6.1 and Figure 6.2 show the initial results. Each square in the confusion matrix shows the number of times that a predicted label correctly matches the text. Additionally, the prediction table shows how well the model did in classifying each category and overall. Based on the confusion matrix, there are a few predicted labels that are never matched up with any of the true labels, such as surgery. Additionally, there is overlap between specialties and other categories like Emergency Room Reports, Discharge Summary, and Notes. Due to the number of categories, the data has also become very sparse, making the results more difficult to read. The prediction results table reflects many of these issues as well. There are categories that have no precision, recall, or f1-score, which shouldn't happen.

In order to fix these issues, domain expertise, SMOTE, and narrowing down the categories were used to improve the results. SMOTE was used to generate additional minority class samples to address the issue of data imbalance because some classes are minorities. Domain expertise and condensing the categories were used to solve the issue of overlap between specialties and categories, along with helping to make the data less sparse. Figure 6.3 and 6.6 shows the results after applying these methods. The features have improved precision, recall, and the f1-score after limiting the categories, as seen in the visualizations. Many of the issues that hurt the initial results are now fixed. Data is less sparse and far more accurate. We have used scispacy models to detect medical entities in our text classification. After a reduction in categories, the feature ophthalmology took the position of the initial maximum accuracy, which was for the feature pain management. The precision

is the highest for ophthalmology, while the maximum recall for surgery has been altered to psychology.

There are various reasons why logistic regression might be a viable choice for healthcare text analysis. It is a well-established method that scholars and practitioners use and understand. It can deal with high-dimensional data, such as text data with a large number of different words or attributes. It is frequently used in healthcare text analysis because it is simple to implement, easy to comprehend and can handle high-dimensional data, such as text data with a large number of different words or attributes. In this work, we developed ML-based Health Care text analysis models for disease classification.

- Our models could be useful in health sectors to classify diseases from healthcare data without the intervention of humans.
- Our models are faster and more accurate as compared to human classification. While evaluating data on diagnoses and procedures, these categorization techniques may be applied in various ways.
- The mutually exclusive categorization approach should be useful for people who want to aggregate their data into a reasonably small number of diagnostic and procedure categories for reporting purposes. Using this strategy, disease-prone populations can be identified.
- The classifications have been expanded to provide statistical information on reasonably specific circumstances.
- Medical Cloud on AWS platform can be used to archive healthcare records.
- Auto-classify medical transcripts based of medical entities.
- Better insights of patient medical history for faster medical treatment.

Future Work

We can leverage AWS Cloud services for healthcare industry to build an robust IT solution to archive the medical prescriptions and analyze it using ML services. Additionally, we can fine tune the healthcare text classification using other ML techniniques like distilbert-multi-class-text-classification and tensorflow. In case of imbalanced datasets, we can leverage it by assigning weights to the target labels to get better results. Additional research and refinement of NLP models, including transformer-based architectures (BERT, GPT-3, etc.) designed with healthcare text data in mind. These models have the potential to enhance entity recognition, information extraction, and comprehension of medical terminology in the context of healthcare literature. Addressing these issues will probably result in more ethical, accurate, and efficient applications of healthcare text analytics as the field develops, which will enhance patient care, diagnostic precision, and overall healthcare outcomes.

BIBLIOGRAPHY

[1] "Exploring knowledge transfer in clinical natural language processing tasks." [Online]. Available: https://web.stanford.edu/class/cs224n/reports/final_summaries/summary031.html

[2] Y. Zhao and H. Che, "SkIn: Skimming-Intensive Long-Text Classification Using BERT for Medical Corpus," Sep. 2022, arXiv:2209.05741 [cs]. [Online]. Available: http://arxiv.org/abs/2209.05741

[3] S. K. Prabhakar and D.-O. Won, "Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention," *Computational Intelligence and Neuroscience*, vol. 2021, p. e9425655, Sep. 2021. [Online]. Available: https://www.hindawi.com/journals/cin/2021/9425655/

[4] L. Yao, C. Mao, and Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC Medical Informatics and Decision Making*, vol. 19, no. 3, p. 71, Apr. 2019. [Online]. Available: https://doi.org/10.1186/s12911-019-0781-4

[5] G. Ivanov, "Mining and Classifying Medical Documents," May 2021. [Online]. Available: https://towardsdatascience.com/mining-and-classifying-medical-text-documents-1876462f73bc

[6] B. Percha, "Modern Clinical Text Mining: A Guide and Review," *Annual Review of Biomedical Data Science*, vol. 4, no. 1, pp. 165–187, Jul. 2021. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-030421-030931

[7] N. S. Pagad, P. N, K. K. Almuzaini, M. Maheshwari, D. Gangodkar, P. Shukla, and M. Alhassan, "Clinical Text Data Categorization and Feature Extraction Using Medical-Fissure Algorithm and Neg-Seq Algorithm," *Computational Intelligence and Neuroscience*, vol. 2022, p. 5759521, Mar. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8920702/

[8] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046417302563

[9] N. Vielfaure, "Medical records redefined: the value of the archival record in medical research," 2015. [Online]. Available: http://hdl.handle.net/1993/30727

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302

[11] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Information Sciences*, vol. 565, pp. 438–455, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521002863

[12] "Clinical Text Classification." [Online]. Available: https://kaggle.com/code/ritheshsreenivasan/clinical-text-classification

[13] M. U. Bokhari, Q. M. Shallal, and Y. K. Tamandani, "Cloud computing service models: A comparative study," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 890–895. [Online]. Available: https://ieeexplore.ieee.org/document/7724392

[14] "Aws Textract." [Online]. Available: https://docs.aws.amazon.com/textract/latest/dg/what-is.html

[15] B. Guzman, MS, I. Metzger, MS, Y. Aphinyanaphongs, M. D., P. D., H. Grover, and P. D, "Assessment of Amazon Comprehend Medical: Medication Information Extraction," Feb. 2020, arXiv:2002.00481 [cs]. [Online]. Available: http://arxiv.org/abs/2002.00481

[16] "Aws S3." [Online]. Available: https://aws.amazon.com/pm/serv-s3/

[17] "Aws Lambda." [Online]. Available: https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

[18] "Aws Comprehend." [Online]. Available: https://docs.aws.amazon.com/comprehend/latest/dg/what-is.html

[19] "Aws Comprehend Medical." [Online]. Available: https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-welcome.html

[20] "Amazon DynamoDB." [Online]. Available: https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Introduction.html

[21] P. Bhatia, B. Celikkaya, M. Khalilia, and S. Senthivel, "Comprehend Medical: A Named Entity Recognition and Relationship Extraction Web Service," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 1844–1851.

[22] "Transcribed Medical Transcription Sample Reports and Examples - MTSamples." [Online]. Available: https://mtsamples.com/