SUMMER TRAINING

At

Public Health Foundation of India, Gurugram

May 19- July 9, 2021

Under the Guidance of Dr. Nutan Jain

A Report By

Shivam Singh Teotia

Roll no-1199

MBA - HEALTH MANAGEMENT- 2020-2022

Institute of Health Management Research, Jaipur

1, Prabhu Dayal Marg, Near Sanganer Airport, Jaipur, Rajasthan 302029

भारतीय
जन स्वास्थ्य
प्रतिष्ठान

PUBLIC
HEALTH
FOUNDATION
of INDIA

Ref: T310/23/07/2021/eCERT

23 July 2021

To whomsoever it may concern

This is to certify that Mr Shivam Singh Teotia has successfully completed his internship from 19 May 2021 to 23 July 2021 with Public Health Foundation of India (PHFI). During his internship he worked under project "Estimation of community level cause specific mortality using in-hospital deaths at selected sites in India".

He was found to be hardworking and sincere during this period. We wish him the very best in all his future endeavours.

For **Public Health Foundation of India**

Aparajita Roy
Director– Human Resources

OFFICE ADDRESS
Plot No. 47, Sector 44,
Institutional Area,
Gurugram 122 002, India
P +91 124 478 1400
F +91 124 478 1601
E contact@phfi.org

REGISTERED OFFICE
431A, 4th Floor Rectangle No.1
Behind Saket Sheraton Hotel
Commercial Complex D4, Saket
New Delhi 110 017
P +91 11 665 440 48

IIPH CENTERS
Gandhinagar
Hyderabad
Delhi
Bhubaneswar
Shillong
Bengaluru

WWW.PHFI.ORG

Table of Content

ACKNOWLEDGMENT

A successful project is a combination of our efforts, encouragement, guidance from experienced people., express my acknowledge and extend heartfelt recognition to the following individuals, who make the project possible.

I express my sincere thanks to IIHMR Jaipur, which gave me this opportunity. My gratitude to

Dr. Nutan Jain for his content guidance and education throughout the internship.

My special thanks to Dr. Ashish Awasthi (Project Head at PHFI) for his co-operation, help, and encouragement.

I sincerely thank my organization mentor Dr. Pooja C Dongare (Fellow Researcher) for her constant guidance throughout the summer training.

List of Abbreviations

AI- Artificial Intelligence

CNN- Convolutional Neural Network

EHRs- Electronic Hospital Records

ICD- International Classification of Disease

ML- Machine Learning

NER- Named Entity Recognition

NLP- Natural Language Processing

NLTK-Natural Language ToolKit

PHFI- Public Health Foundation of India

TL- Transferred Learning

TSV- Tab Separated Value

WHO- World Health Organisation

At the end of the first-year of MBA HM at IIHMR University, students are supposed to go for summer training at any health sector organization for two-month period. The trainee joined the Public Health Foundation of India (PHFI), Gurugram for summer training but worked from home due to COVID 19 infection From May 19, 2021 the trainee worked under the guidance the mentor at PHFI. The trainee was assigned responsibility to build an algorithm to extract the disease names from the patient's death summary, which was provided by the hospital in the pdf form. The learning was how natural language processing works and name tagging to the text files.

A Snapshot of Activities

Duration	Activities
May 19-28, 2021	Learned about NLP, ENR, and Spacy
May 31-4, 2021	Collecting Data for training model
June 7-18, 2021	Start training of the NER model
June 21-25, 2021	Clearing noise from the text file
June 28-2, 2021	Start testing model on text files
June 5-9, 2021	Retrained the model using new data set

ABOUT THE ORGANISATION

The Public Health Foundation of India (PHFI) is a public-private initiative that has collaboratively evolved through consultations with multiple constituencies including Indian and international academia, state and central governments, multi & bi-lateral agencies, and civil society groups. PHFI is a response to redress the limited institutional capacity in India for strengthening training, research, and policy development in the area of Public Health.

The Prime Minister of India, Dr. Manmohan Singh, launched PHFI on March 28, 2006, in New Delhi. It Structured as an independent foundation, PHFI adopts a broad, integrative approach to public health, tailoring its endeavor to Indian conditions and bearing relevance to countries facing similar challenges and concerns. The PHFI focuses on broad dimensions of public health that encompass promotive, preventive, and therapeutic services, many of which are frequently lost sight of in policy planning as well as in popular understanding. PHFI recognizes the fact that meeting the shortfall of health professionals is imperative to a sustained and holistic response to the public health concerns in the country which in turn requires health care to be addressed not only from the scientific perspective of what works but also from the social perspective of, who needs it the most.

PHFI has also built up an impressive portfolio of research and implementation projects, funded by reputed national and international agencies through competitive grants. With over 1600 publications in scientific journals and an average impact factor of 5.3, PHFI has established a creditable track record in research and has been so recognized by the Department of Scientific and Industrial Research. More importantly, the research is providing useful inputs to India's health policy and programs in many areas of public health importance. Four funded centers of excellence in chronic diseases, disabilities, equity, and social determinants, and environmental

health are leading applied research projects and capacity development in those areas. At the global level too, faculty and researchers are actively contributing to many initiatives, expert groups, and commissions such as Agriculture and Food Systems for Nutrition, Global Burden of Disease Study, WHO Commission on Ending Child Obesity, Lancet Commissions on Health Professional Education, Mental Health, Investing in Health, Palliative Care and Obesity as well global panels on Antibiotic Resistance. International conferences have been convened by PHFI on maternal health, antibiotic resistance, the endgame for tobacco, global youth meet on health, health in sustainable development, and new directions for public health education.

Vision of the Organization

“Our vision is to strengthen india’s public health institutional and systems capacity and provide knowledge to achieve better health outcomes for all.”

Mission of the Organization

Developing the public health workforce and setting standards. Advancing public health research and technology. Strengthening knowledge application and evidence-informed public practice and policy.

PROJECT REPORT

ESTIMATION OF COMMUNITY-LEVEL CAUSE-SPECIFIC MORTALITY USING IN-HOSPITAL DEATHS AT SELECTED SITES IN INDIA

INTRODUCTION

According to the RBD Act 1969, it is mandatory to mention the cause of death on the death certificate. The medical certification of cause of death is a predominant Medico-legal document. All the medical practitioner has the legal responsibility to ensure the rightly mention the cause of death. Medical Certification of Cause of Death (MCCD) is a Government scheme under the RBD Act 1969.

In India, the death certification is believed to be very poor in most hospitals. A Hospital-based retrospective study on assessing the accuracy in completing the Medical certification of cause of death was conducted at a rural Medical college teaching Hospital in Andhra Pradesh for one year from 1st January 2011 to 31st December 2011. About 110 MCCD forms along with case sheets are collected from the MRD section are examined for this study. And in the study, it was found that in any of the certificates the certifier didn't mention the ICD codes of the disease. WHO always insists that every MCCD, should mention the ICD 10 codes of the disease.

During the COVID pandemic, patients load on every hospital is increased unexpectedly and the human resources were limited, all the doctors were busy in saving the life of patients and making the impossible possible by putting their own life at risk. It is impossible for them to mention the exact cause of death on the certificates and also assigning the ICD-10 codes with respect to the disease.

For easing the work of medical practitioners and also minimizing the chance of human errors, our team taking the help of Artificial Intelligence (AI). Many new technologies emerge in the digital age leading to the development of Artificial Intelligence. Machine learning is one of the outstanding fields of AI that reduces workloads and enhances data from large amounts of data. The role of AI in various fields such as education financial management, transportation, health management, etc. It generates many opportunities for AI investigators and skilled developers. Health management is part of the Life science field that has emerged with great strides in AI technology. In the case of Covid-19, AI played a major role in determining the diagnosis Researchers detected the disease by training data sets obtained using a machine learning model and were able to predict the presence of Covid-19. However, it is important to increase the overall accuracy of the machine learning model.

On the other hand, a large number of health-related articles have been published in popular archives such as PubMed, WebMD, etc. These articles have a lot of hidden information. Many organizations and drug manufacturers have invested in extracting hidden information from large amounts of information.

Pharmacovigilance is one of the major areas, where, you have to monitor the label of a drug before it can be released into the market. Extracting annotations such as crno, disease, chemical compound, gender, symptoms rate, date, location, etc. It is hard work. The vast majority of textbooks available online are in an unstructured format, for example, electronic health records, clinical trials have a variety of methods depending on internationa

law. Identifying the aforementioned businesses in a random document is a challenging task for researchers and developers. Making a toxicology report is highly dependent on these organizations, extracting detailed studies and other information. Several researchers also report on the work of generators to manually extract and produce clinical research reports, EHRs simplify the ML model and reduce their time and dependence. Most documents are available in a non-standard format and it is a man's job to access all the information from the documents. Many methods such as dictionary-based, learning-based, and hybrid-based methods are proposed to exclude these organizations from a large number of informal data. The method based on the dictionary is much more economical compared to the method based on machine learning. Maintaining and restoring corpora leads to significant hardware and storage costs. The machine-based method of continuous transformation into Transfer Learning (TL) produces more precision.

Building a TL model is being easy by the recent progression of the database technologies which easy to construct a huge amount of corpora. The corpora can be built from the information available on the pharmaceutical web pages. This information can be used for the various Natural Language Processing (NLP) applications to get wisdom. NLP undertaking these challenges and provide valuable insights. This emerging technology needs life science dictionaries and a working machine learning model to predict the named entities. This report has been concentrated on Named Entities Recognition (NER) on the patient's death summary generated from the hospitals while discharging the patient's body. A custom based approach has been selected to identify the entities. Diseases, symptoms, dosage forms, and route of administration. Separate dictionaries have been built for the above-mentioned entities. The unstructured documents are converted into a structured format and the named entities are matched with the string in dictionaries. The matched words are annotated with the beginning and end position of the characters. The annotated sentences are trained with the Spacy model and a custom spacy model was built. The model was evaluated with the dictionary and the confusion matrix is calculated.

Our team get the files of patients' death summary in pdf format from the hospital's EHRs department and then converted into a text document and then running our Natural Language Processing model over the text documents to extract the medical terminologies from it and then assign the ICD-10 codes to every patient death summary files.

The International Classification of Disease (ICD) is a standard diagnostic tool developed by the World Health Organization (WHO), to monitor the incidence and spread of diseases and related conditions.

The ICD has a wide range of clinical applications and is used not only by doctors but also by emergency workers, insurance companies, researchers, and policymakers. The ICD is used to diagnose diseases and keep diagnostic data for therapeutic, quality, and disease purposes as well as reimbursement for insurance claims.

The value of the ICD-10 code system can be assessed through its use in various areas of quality management, health care, information technology, and public health.

- The ICD-10 code system provides accurate and up-to-date procedures to improve health care costs and ensure appropriate refund policies. The current codes help health care providers identify patients in need of immediate disease management and plan effective disease management programs.

- ICD-10-CM is internationally recognized to facilitate the implementation of quality health care and its comparability globally.
- Compared to the previous version (ICD-9-CM) ICD-10-CM is more specific and treats public health diseases, particularly external injury-related diseases, e.g. terrorism.
- ICD-10 codes hold some value in research because code analysis is an important part of research and development. The coding and logic system allows a few coding errors to ultimately benefit from research and development analysis.
- The improved type of ICD code system enhances health policy decision-making by providing better details of monitoring and performance of the organization.
- The ICD-10 code system is much simpler and reversible into an electronic format that offers a better format than ICD-9, other codes such as SNOMED CT and CPT codes.
- ICD-10 codes are specifically designed for refund purposes to provide a sound basis for the payment process.
- Alphanumeric formats for the ICD-10 code system provide a better alternative than ICD-9-CM codes that offer a more flexible and scalable type e.g. diabetes mellitus - E10-E14
- Finally, the ICD-10 encoding system helps to:
    - Minimize medication error
    - Improving treatment options and disease outcomes
    - Low medical costs and claim costs
    - In health policy and operational planning and strategies
    - Improving payment systems by processing claims
    - Limit claims submission
- ICD-10 codes shape the future of the medical practice. The sooner doctors start by involving them in managing their practice, the better for them and their patients.

Certification of causes of deaths help in decision making for better health planning, developing policies, cause of deaths statistics, and also for the life insurance payments. With the help of AI, recognition of disease-named entities automatically from EHRs can be done without any human interference. It is very difficult to assign ICD codes to the patient's death certificate by hand because hundreds of daily reports are made by hospitals and this task requires a lot of human effort. Therefore, by building an algorithm in machine learning it will assign ICD codes to files concerning diseases

Study Objectives

The objective of the study was to estimate cause-specific mortality at the community level using the available in-hospital data at selected sites in India and develop an algorithm to assign ICD-10 codes to in-hospital mortality cases.

3.1. Literature Review

1) **AUTHOR**- Jasow Fries

   Year of published-2017

   Title-" A generative model for biomedical named identity recognition without labelled data"

   Providing unlabelled documents and defining weak supervision input. Using generators to convert documents into a collection of entities for classification. Autogenerating labeling functions using structured resources or user heuristics.Fitting a multinomial generative model using the output of all labeling functions as applied to candidates; and Generating probabilistically labeled data.

   This study approach intrinsically to automatically construct large training sets, allowing SWELLSHARK to teach high-performance taggers using state of recent deep learning models.

2) **Author**- Luqi Li and Jie Zhao

   Year of Published-2019

   Title- An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records

   Firstly, some pre-processing steps including data annotation, sentence splitting, and character segmentation were performed.

   Secondly, additional features including dictionary and part-of-speech were introduced into this study. dictionaries supported the terms from three categories ("Body structure", "Procedure", "Pharmaceutical/biologic product") associated with NER task.

   And third, the attention mechanism can form up for the shortcomings of the semantic representation within the traditional encoder-decoder model, and it's easier to capture the long-distance interdependent features in the sentence.

3.2. METHODOLOGY

- The task to find the best machine learning tool for extracting desired information from text files.
- The task to compare the different algorithms and select the best of them.
- The task to do a study on the selected algorithm for the project and do a crash course on it.
- The task to collect data from the different medical sites for model training.
- The task to convert the available data into spacy format for training purposes.
- The task to develop the convolutional neural network training model in spacy and build a model.
- The task to do the data cleaning of patient's death summary files.
- The task to extract disease names from text files using a trained model.

The following steps were taken to conduct the study

3.2.1. Step-1

The first task given to me in the first week of my internship was to find out which machine learning tools for NLP are used to extract the useful information from the text and also find out how that particular tool will work. For this task I go through many websites for collecting the useful information and selected exzctely 5 heroic tools for NLP.

3.2.2. Step-2

In the second week of my internship, I had to report my mentor with the collected data and also suggested her the best choice of tool for our research work. So, I had collected 5 heroic NLP tools for the project and here is the list of all the tools with data for comparison.

- Core NLP, from Stanford group
- NLTK, the most widely-mentioned NLP library for Python
- TextBlob, a user-friendly and intuitive NLTK interface
- Genism, a library for document similarity analysis
- Spacy, an industrial-strength NLP library built for performance

From the above mentioned Python libraries for NLP, I used two of them in our project. I used NLTK for Text cleaning and Spacy for NER model training.

3.2.3. Step-3

In the third week of my internship,I contribute my time for understanding the working of selected Python libraries. I did a crash course on Spacy over youtube and got to know, how this Python library works and how I can utilized it in our project work. There are many function which are not required for our project and it's the importank thing to find out what is important and which thing will waste our time. So, this crash course was completed in 1 week and I was completely ready to start working on the given project.

3.2.4. Step-4

I am working on a project in which, the main work is to extract the disease name from the patient's death summary file, and for this particular task I need the data set of all the diseasse name including in it. So, I again start my search for this particular data set and my search endup on github profile of BioBERT. On this BioBERT github profile I found the 2 folders of containing all the disease names in the format of TSV(Tab seperated Values). The data which I got from the github was present in the format of TSV and my next task is to convert this format into Spacy specific format. In these TSV files, each term is defined using the BIO format. A few lines from train.tsv in BC5CDR-databases look like this:

ventricular **B**

tachycardia **I**

during **O**

low **O**

dose **O**

intermittent **O**

dobutamine **O**

treatment **O**

in **O**

a **O**

B-> Begin entity,

I-> inside entity

O-> outside entity

3.2.5. Step-5

After downloading the data from the github, now its time to convert it into Spacy specific format. This is the crucial phase of the proposed NER model. Preliminary consideration is the heart of the proposed structure that clears raw data and provides a detailed sentence with entities. This step involves building unstructured data into a profitable format. TSV format has been used for this model to save the train and test data. The raw data are tokenized by annotating the sentences with the entity. For training, model annotated sentences are essential. The customized database was developed within the three entities: Begin entity(B), Inside entity(I), Other entity (O). Descriptive sentences contain the beginning and end of the characters. This makes it easier to train the model.


```
[["Selegiline - induced postural hypotension in Parkinson ' s disease : a longitudinal study on the effects of drug with
drawal .", {'entities': [(21, 29, 'B_Disease'), (30, 41, 'I_Disease'), (45, 54, 'B_Disease'), (55, 56, 'I_Disease'), (5
7, 58, 'I_Disease'), (59, 66, 'I_Disease')]}], ["OBJECTIVES : The United Kingdom Parkinson ' s Disease Research Group (
UKPDRG ) trial found an increased mortality in patients with Parkinson ' s disease ( PD ) randomized to receive 10 mg se
legiline per day and L - dopa compared with those taking L - dopa alone .", {'entities': [(32, 41, 'B_Disease'), (42, 4
3, 'I_Disease'), (44, 45, 'I_Disease'), (46, 53, 'I_Disease'), (132, 141, 'B_Disease'), (142, 143, 'I_Disease'), (144, 1
45, 'I_Disease'), (146, 153, 'I_Disease'), (156, 158, 'B_Disease')]}], ['Recently , we found that therapy with selegilin
e and L - dopa was associated with selective systolic orthostatic hypotension which was abolished by withdrawal of seleg
iline .', {'entities': [(92, 100, 'B_Disease'), (101, 112, 'I_Disease'), (113, 124, 'I_Disease')]}], ['The aims of this
study were to confirm our previous findings in a separate cohort of patients and to determine the time course of the car
diovascular consequences of stopping selegiline in the expectation that this might shed light on the mechanisms by which
the drug causes orthostatic hypotension .', {'entities': [(274, 285, 'B_Disease'), (286, 297, 'I_Disease')]}], ['METHODS
: The cardiovascular responses to standing and head - up tilt were studied repeatedly in PD patients receiving selegilin
e and as the drug was withdrawn .', {'entities': [(97, 99, 'B_Disease')]}], ['RESULTS : Head - up tilt caused systolic o
rthostatic hypotension which was marked in six of 20 PD patients on selegiline , one of whom lost consciousness with unr
ecordable blood pressures .', {'entities': [(32, 40, 'B_Disease'), (41, 52, 'I_Disease'), (53, 64, 'I_Disease'), (95, 9
7, 'B_Disease')]}], ['A lesser degree of orthostatic hypotension occurred with standing .', {'entities': [(19, 30, 'B_Di
sease'), (31, 42, 'I_Disease')]}], ['Orthostatic hypotension was ameliorated 4 days after withdrawal of selegiline and t
otally abolished 7 days after discontinuation of the drug .', {'entities': [(0, 11, 'B_Disease'), (12, 23, 'I_Diseas
```


Fig. 2 After converting TSV to Spacy format for training

3.2.6. Step-6

Now the most important part of my project is start from this stage because if I correctly design the Custom-NER model then my previous work on this project will be useful. The two models were used for extracting the disease name from text files.

METHOD 1: Blank Spacy Model

In the first method, our team is using a custom-NER spacy blank model in the English language from the scratch for training purpose and with our own annotated training data set, which we downloaded from the github profile of BioBERT . The generated database is divided into training and testing databases, 80% of the database is considered a training database and 20% is considered a testing database. After the pre-processing step, the annotated data is transferred to the Custom-NER model. This section trains the model with defined sentences. The model is based on the Convolutional Neural Network (CNN). A blank Spacy model has been used and trained by the newly defined database. The detailed flow chart for the proposed Custom-NER is presented in Fig-1.The proposed NER method was tested using a confusion matrix.

The performance of the proposed hybrid method was assessed with standard measures of precision, recall, and F-score.

TP = Word predicted as either I_Disease or B_Disease and present in the data(train/test/validation) as either I_Disease or B_Disease.

FP = Word predicted as either I_Disease or B_Disease and not present in the data (train/test/validation) as either I_Disease or B_Disease.

FN = Word present in the data(train/test/validation) data as either I_Disease or B_Disease but not predicted as either I_Disease or B_Disease.

Precision

It is a classification of identified businesses, related to the need for user information. An accurate calculation is given in equation 1.

Pr = TP/(TP + FP)

Recall

It is part of the number of businesses identified in question that has been successfully identified. The recall rate for memory is given in Eq. (2).

Recall= TP(TP + FN)

F-points

It is a measure that includes clarity and recall. It is a harmonic definition of accuracy and memory, the traditional F value is given in Eq. (3).

$F = 2 * (precision * recall)/(precision + recall)$

First I trained my model with only 500 iterations and the loss(Lower the loss, better the model training) was around 300, which is not acceptable. Then again I put the iteration on 2000 and the loss was around 120, which is also not acceptable. For the next training my mentor suggested me to put it 5000, then I put the iteration count as 5000 and the whole training of the model was completed in 60 hours and this time loss count was 32 only.

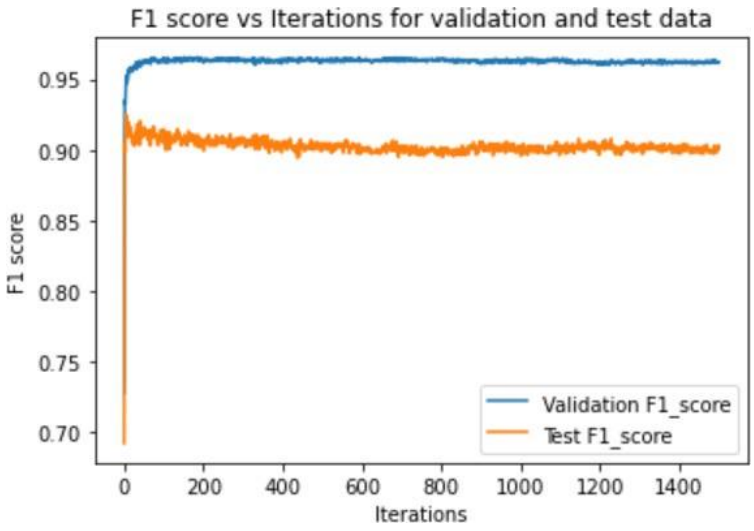

Fig-2. Shows the F-1 Score over the 90%

METHOD 2: Scispacy pre-trained model

In the second method, we used the pre-existing scispacy model (en_ner_bc5cdr_md) trained on BC5CDR corpus for recognizing two entities (Disease and chemical) from the medical text. The BC5CDR corpus consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease interactions. As the model was already trained on medical data, we simply download it from the git-hub and install the model in our system, and applied the trained model on our patient's death summary files.

3.2.7. Step-7

After building the Spacy model, now the next step is to give the patient's death summary files folder as an input to the trained model. But before this step I need to clean the text files because these text files contains lot of noised words. I got the patient's death summary in pdf format from the project head and then I convert the pdf files into a text document using the python library. These files contain the hospital name, location of that hospital, CRNO of the patient, age, gender, diagnosis summary, and cause of death.

Diagnosis: NHL in remisson Malabsorption syndome Small intestinal bacterial Overgrowth Distal Jejunal stricture Severe Malnutrition Shock Renal failure
History 1: K/C/O NHL in remission , bone marrow supression - improved. Now having small bowel diarrhea since 1 yr with significant wt loss. H/o generalise
weakenss & pedal edema.Evaluated, diagnosed to have SIBO. Small bowel enema revealed distal jejunal stricture. started on Rifaximin. Now has come with
increased frequency of stools x 3 days 10 to 15 stools per day watery no blood/ mucus No h/o pain abdomen, fever h/o decreased urine output h/o
altered sensorium+ O/E Drowsy and irritable Pallor+ Glossitis+ Pedal edema + BP - 80/60 PR - 120/mt P/A Soft, no organomegaly, BS + CVS & Chest NAD
Hospital Course: Pt was admitted and found to be severely dehydrated. He was having azotemia and hypokalemia. He was started on IVF, iv inotropes,
antibiotics, TPN. He was planned for central line. But in the meantime he was having severely coagulopathy. So it was withheld for timebeing. But he
continued to worsen and hypotension became refractory to inotropes and he had a cardiac arrest at 11 AM. CPR was done according to protocol. But he
could not be revived and he was declared dead at 11.10 AM on 4/8/09
FUTUREPLAN:
Advice :
*
Investigations :
Investigations1 :

Cause of Death :Severe Malnutrition Shock Renal failure

Fig no 3 is the file that is provided by the hospital EHR department.

DATA CLEANING

After getting the patient death summary file, I took the following steps for cleaning the files. So, data cleaning can be done in 4-5 steps and python has its library to clean the data which is known as NLTK.

- **Deleting additional spaces**

Most of the time text data may contain additional spaces between words, after or before a sentence. So, to begin with I remove these extra spaces in each sentence using regular expressions.

- **Removes punctuation**

The punctuation marks in the text do not add value to the data. Punctuation marks, if they are attached to any word, will create a problem separating other words like(,,.,',/,@,!)

- **Words Normalization**

In this case, I simply turn the case for all the characters in the text, big or small. As python is a sensitive case language so it will treat nlp and NLP differently. One can easily turn the word up or down by using:

str.lower () or str.upper ().

- **Tokenization**

Divide a sentence into words and build a list, which means each sentence is a list of words.

- **Deleting stop-words**

Terms include Me, he, she, and, however, there were, presence, and so on, which do not include meaning in the data. Therefore, these terms should be removed that help reduces features from our data. This is removed after token insertion.

- **Lemmatization & Stemming**

Stemming: The process that takes a word to put its roots. It simply removes the suffixes from the words. The title may not be part of the dictionary, which means it will not provide the meaning. There are two main types of stem-Porter Stemmer and Snow Ball Stemmer (advanced version of Porter Stemmer).

Lemmatization: Move the name to the root form called Lemma. It helps to bring the words to their dictionary form. Automatically applied to nouns. It is more accurate as it uses detailed analysis to create word groups with the same meaning in context, so it is more complex and takes longer. This is used when we need to store content details.

3.2.8. Step-8

In this final step, after cleaning the patient's death summary files using the NLTK libraries, now its time to check the output of our 2 different models,

RESULTS

The output results for the two models are as follows:

OUTPUT RESULT of METHOD 1

After giving the input file to the Custom-NER model the output file contains the crno of the patient, age, and gender of the patient and contains disease names.

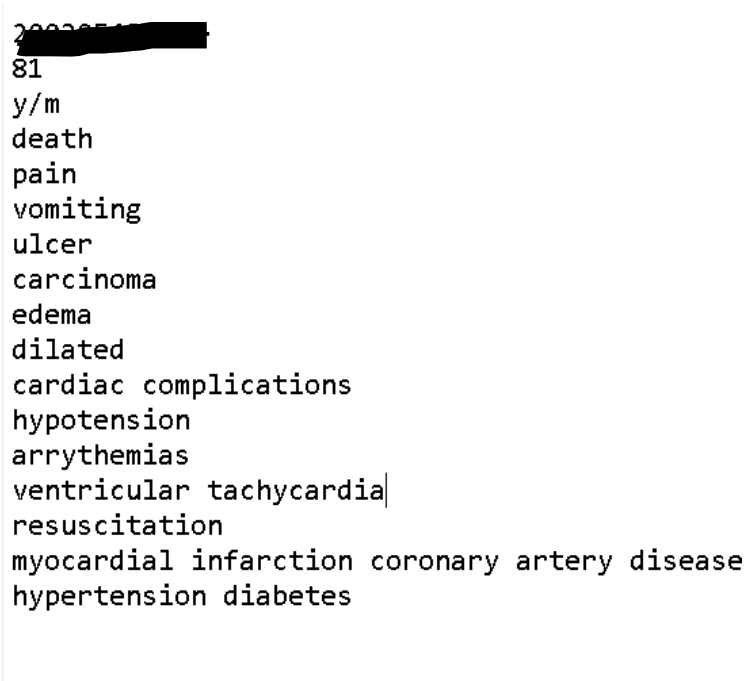

FIG. 4 Output result

OUTPUT FILE of METHOD 2:

After giving the input file to the pre-trained model(bc5cdr) the output file contains the crno of the patient, age, and gender of the patient and contain disease names.

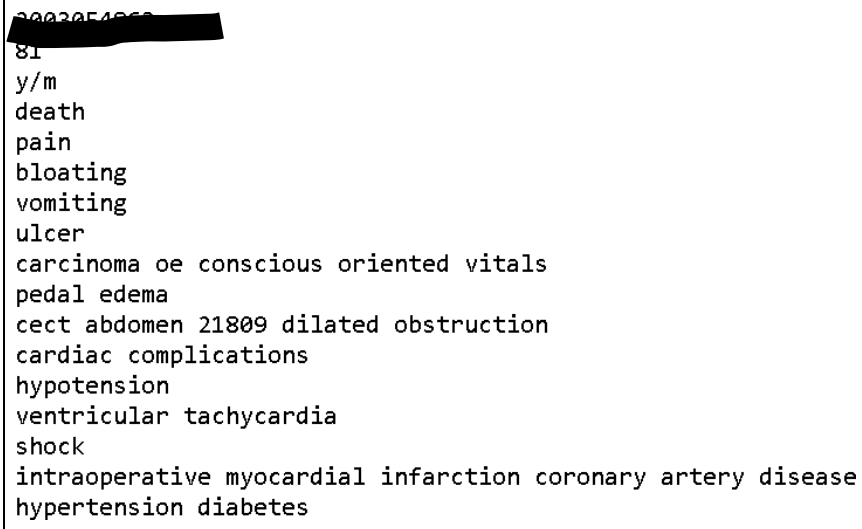

FIG -5. Output Result

Both the models are able to extract the disease names from the patient's death summary files efficiently, but the result of model 2(bc5cdr) also contains more skewed data, which is not acceptable as this skewed data will affect the end result of our project.

On comparison the outputs of both the model, Custom-NER contains less noise data as compare to Bc5cdr model. There are some words in the result of bc5cdr model like bloating, dilated, obstruction, shock which are not required and being treat as skewed data.

DISCUSSION

Most of the time and it's also found in different studies that hospitals don't follow a proper set of rules in making death certificates and most of the time in government hospitals, doctors usually forget or skip mentioning the ICD codes. This project is solving this kind of problem and designing an autogenerated ICD-10 codes for the death certificates.

In this study, our team works on different algorithms and compares these algorithms with already existed algorithms. Our team observed that the pre-existing models are trained on very limited data and also failed to recognize some disease names in patients' death summary files and many a time this pre-existed model also extracts skewed data from the text files, which has no relevance for our work.

A key observation of the results presented is that our Custom-NER model successfully extract all the disease names from the patient's death summary files and it outperformed the already existed sicspacy model which is trained on a very limited training data set.

CONCLUSION

The project under I was assign is to find the extimation of commuinty-level cause-specific mortality using in-hospital deaths at selected sites in India. And the task assign to me is extract the disease name from the patients death summary. For the assign task I worked on Natural language processing to extract disease names. During the internship I worked on two methods to extract entities from text. Scispacy model is already existed and trained with annoteated dataset develop by field experts, and the second model is design by our team by using the CNN and we named it Custom-NER model. When our custom-NER model was compared with scispacy model, transmission readings show a 5% increase in accuracy. Our model outperformed already existed model. The proposed Custom-NER model were trained on 80% of training dataset and 20% of testing dataset, respectively. The most significant aspect of the project was gathering all documents containing disease names to create a well-maintained training and testing dataset. The tested model's F1 rating has been steadily improving and the loss was around 32, which is a good number. We intend to increase the number of entities and retrain the model with the new entities in the future to improve the model's overall accuracy. More accuracy can be gained by enriching dictionaries with more accurate data.

REFERENCES

[1] Public Health Foundation of India [Internet]. Public Health Foundation of India -. 2021 [cited 8 July 2021]. Available from: https://phfi.org/

[2] The Practo Blog for Doctors. 2021. *What are ICD-10 codes and why are they important for doctors? - The Practo Blog for Doctors*. [online] Available at: <https://doctors.practo.com/icd-10-codes-important-doctors/> [Accessed 27 June 2021]

[3] GitHub. 2021. *dmis-lab/biobert*. [online] Available at: <https://github.com/dmis-lab/biobert> [Accessed 6 July 2021].[3] Spacy.io. 2021. [online] Available at: <https://spacy.io/> [Accessed 27 June 2021].

[4] Talix. 2021. *Talix Platform | Unlock the Power of Unstructured EHR Data*. [online] Available at: <https://www.talix.com/platform/> [Accessed 28 June 2021].