Summer Training Report

MENTOR-DR. NUTAN JAIN

BY SHIVAM SINGH TEOTIA

ROLL NO-1199

About Organization

PUBLIC HEALTH FOUNDATION OF INDIA

- Public Health Foundation of India (PHFI) is a public-private initiative
- 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 former Prime Minister of India, Dr. Manmohan Singh, launched PHFI on March 28, 2006, in New Delhi.

Project Title and Objective

DEVELOPMENT OF ALGORITHM IN MACHINE LEARNING FOR FACILITATING CAUSE SPECFIC MORTALITY

- **A Sub Study of the Project entitled**

  Estimation of community-level cause-specific mortality using in-hospital deaths at selected sites in India

- **Objective**

  Objective of the study was to develop AI based algorithm to assign ICD-10 codes to hospital mortality cases

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.

During the COVID pandemic, doctors found difficult to mention the exact cause of death on the certificates and assigning the ICD-10 codes due to lack of time.

For easing the work of medical practitioners and also minimizing the chance of human errors, Artificial Intelligence (AI) can be used for building an algorithm to find out all disease names.

Literature Review

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

   Year of Published-2019

   Title- “Clinical named entity recognition in Chinese electronic medical records using an attention-based deep learning model.”

   To begin, certain pre-processing stages were completed, including data annotation, sentence splitting, and character segmentation.

   Second, new characteristics such as dictionary and part-of-speech were included in this research. The terms from three categories (“Body structure,” “Procedure,” and “Pharmaceutical/biologic product”) linked with the NER work were supported by dictionaries.

   Third, the attention mechanism can compensate for the semantic representation's deficiencies in the standard encoder-decoder paradigm, making it easier to capture the sentence's long-distance interdependent aspects.

Contd…

2) **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 labelling 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 labelled data.

Methodology

Receiving Patients' death summary in pdf format from the hospital's EHRs department and converted into a text document and then running 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 following steps were taken to conduct various tasks

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

Step-1

The task to find the best machine learning algorithms for extracting data from pdf, text files.

The first task that was assign to me in the first week of my internship was to find out the best machine learning tool for extracting desired information from text files. The team had collected 5 heroic NLP tools for the project.

- **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

Step-2

- Understanding the working of selected Python libraries.
- Crash course on Spacy over Youtube to know, how this Python library works and its application to the study.
- Identify useful things to the study.

Step-3

- Extracting the disease name from the patient's death summary file was the main task.
- Collected data set of all the diseases name.
- Found the 2 folders of containing all the disease names in the format of TSV (Tab separated Values) on github profile of BioBERT. In these TSV files, each term was defined using the BIO format. A few lines from train.tsv databases look like this:

Ventricular	**B**	tachycardia	**I**	during	**O**	low	**O**	dose	**O**

Step-4

- After downloading the data from the github, converted it into Spacy specific format.
- This is the crucial phase of the proposed NER model.
- This step involves building unstructured data into a spacy specific format. The raw data were tokenized by annotating the sentences with the entity. For training, model annotated sentences are essential. 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
```

Step-5

❖ **The task to develop the convolutional neural network model by using spacy.**

The team used two methods for extracting disease names,

METHOD 1: Blank Spacy Model

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 model is based on the Convolutional Neural Network (CNN)
The proposed NER method was tested using a confusion matrix

Contd…

- **The task to develop the convolutional neural network model by using spacy.**

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 the system, and applied the trained model on patient's death summary files.

Step-6

❖ **The task to do the data cleaning of patient's death summary files.**

NLTk library for cleaning the files of patients' death summary was used. The whole process was completed in 4-5 steps.

- ❑ **Deleting additional spaces**
- ❑ **Removes punctuation**
- ❑ **Words Normalization**
- ❑ **Tokenization**
- ❑ **Deleting stop-words**
- ❑ **Lemmatization & Stemming**

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.

2003054863
81
y/m
death
pain
vomiting
ulcer
carcinoma
edema
dilated
cardiac complications
hypotension
arrythemias
ventricular tachycardia
resuscitation
myocardial infarction coronary artery disease
hypertension diabetes

Contd…

OUTPUT RESULT 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.

```
2003054863
81
y/m
death
pain
bloating
vomiting
ulcer
carcinoma oe conscious oriented vitals
pedal edema
cect abdomen 21809 dilated obstruction
cardiac complications
hypotension
ventricular tachycardia
shock
intraoperative myocardial infarction coronary artery disease
hypertension diabetes
```

Contd….

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.

Conclusion

- Scispacy model is already existed and trained with annotated dataset develop by field experts.
- The second model was designed by the team by using the CNN, named as Custom-NER model.
- When custom-NER model was compared with Scispacy model, transmission readings show a 5% increase in accuracy. The model outperformed already existed model.
- The proposed Custom-NER model was 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.
- The team 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.

THANK YOU