International Professional Doctorate and Postgraduate Symposium 2021 (iPDOCs2021)
Universiti Teknologi Malaysia, 25 Sept 2021

A Processing Model Using Natural Language Processing (NLP) For Clinical Notes For Producing Symptoms Of COVID-19

01 INTRODUCTION

The novel coronavirus disease, also referred to as COVID-19 pandemic, was first reported in China in December 2019[1]. Because of its pathogenicity and global dissemination, it is currently considered a serious global health concern[3].

In recent times, artificial intelligence (AI) has been considered as a potentially powerful tool in the fight against many evolving pandemics.

Regarding the ongoing COVID-19 pandemic, dozens of research efforts have emerged, and a majority of the articles focused on the importance of harnessing artificial intelligence technologies to diagnose COVID-19 patients. Although much has been learned about the COVID-19 since its discovery, there are still numerous unanswered issues about.

Clinical notes in free-text format give crucial information for answering these concerns. Text mining is an active research area that has attracted significant attention due to the imminent need to transform these data into useful information and knowledge[5]. Therefore, the texts could be processed using a natural language processing (NLP) system to determine the symptoms.

2.2 Text Pre-processing

➔The clinical notes pre-processing module performs COVID-19–related texts cleaning, sentence chunking and parsing tasks to prepare texts for extraction of information components from pure text.

➔ Moreover, we also eliminating useless words from clinical notes, which are defined as stop-words in NLP.

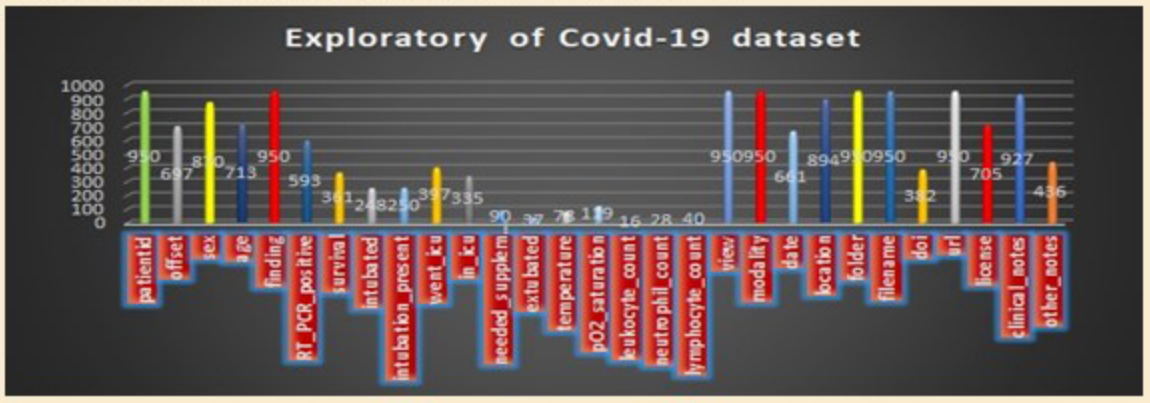
➔ We develop the proposed model and perform the text analysis on a Python programming environment using many Python packages, such as nltk, gensim, and sklearn library.

04 CONCLUSION

We presented an NLP-based system for the automated extraction of COVID-19 symptoms from unstructured clinical notes. Moreover, our findings may aid in improving practical strategies for healthcare services as the clinical decision support systems related to COVID-19. Further investigation in the current and any future coronaviruses and methods for automated generation of data is the subject of future work.

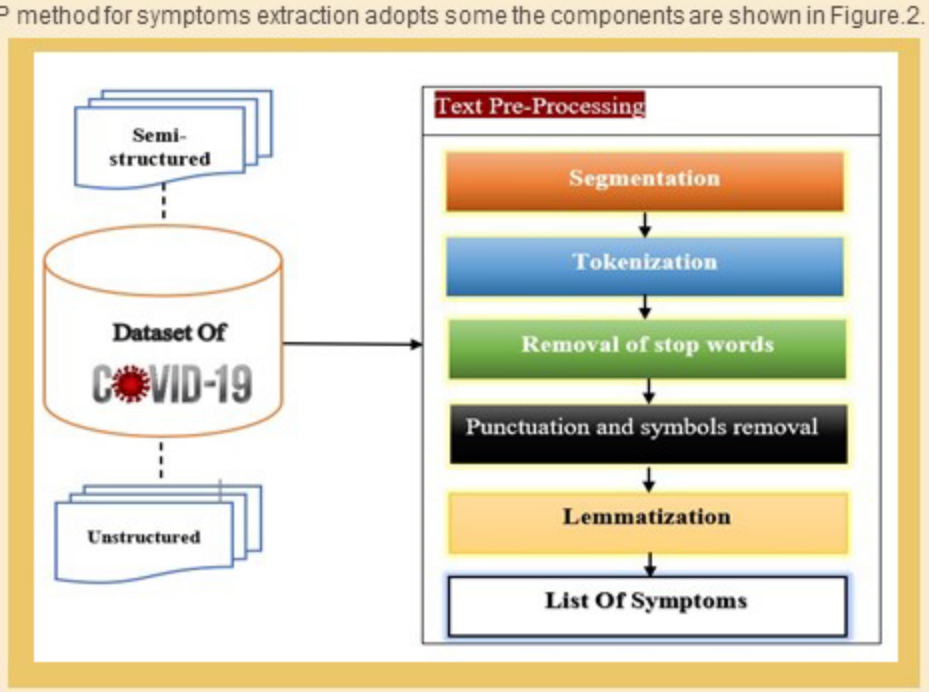
02 MATERIEL: 2.1. Preparing the Clinical data

The Johns Hopkins University (JHU) dataset is one of the most commonly used by researchers and journal media. As a first step in developing this model, this paper focuses on COVID-19–related clinical data. We have collected data from GitHub, an open-source data repository(https://github.com/Akibkhanday/Meta-data-of-Coronavirus). In which about 950 patient's data is stored which have shown symptoms of COVID-19 and other viruses, the graph below shows an exploratory visualization of the data (Figure 1).



03 DISCUSSION AND RESULTS

The NLP method for symptoms extraction adopts some the components are shown in Figure.2.



Based on figure 2, it can be concluded that the overall text processing with NLP starts with dividing entry texts into sentences, followed by other operations as shown in Table 1. Furthermore, we have created a list of 7223 stop words. For example: "60 year old male patient. with sick cough asthenia and fever for three days. Coming from the high-risk zone. RT-PCR test positive. No previous medical history".

Table 1: Example of some NLP stages results

No	Sentences & Remove stope-words	Tokenization
1	60 year old male patient	{"60" "year" "old" "male" "patient" "sick" "cough" "asthenia" "fever" "three" "days" "coming" "high-risk" "zone" "RT-PCR" "test" "positive" "No" "previous" "medical" "history"}.
2	sick cough asthenia fever three days	
3	Coming high-risk zone	
4	RT-PCR test positive	
5	No previous medical history	

Authers Name:
- Amir Yasseen Mahdi (Presenter)
- Prof.Dr.Siti Sophiayati Yuhaniz

Organization Name: UTM/ FTIR
Matric No: PRT193049
Telephone: 01160524752

Authers E-mail:
mahdi.amir@graduate.utm.my
sophia@utm.my