International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

Natural Language Processing and summarization of medical symptomatic data from geographical diverse locations

Snehal N. Palve[1], R.N Awale [2], Vaibhav Awandekar [3], Sunil Lakdawala [4]

[1] *M.Tech Student, Electrical Department, Veermata Jijabai Technological Institute, Mumbai, India*
[2] *Professor, Electrical Department, Veermata Jijabai Technological Institute, Mumbai, India*
[3] *Senior R&D Engineer, A3 Remote Monitoring Technologies Pvt Ltd, India*
[4] *Director, A3 Remote Monitoring Technologies Pvt Ltd, India*

Abstract - *This Paper deals with detection of regional symptomatic diseases at an earlier stage. This early detection across various geographical locations will be helpful for early diagnosis and for death prevention on a larger scale. Adopting random methods for early detection and its respective factors will be unsystematic. The proposed method is for detecting disease epidemics/pandemics by considering large symptomatic data and Natural Language Processing (NLP). These Symptomatic data is available in comments made by the doctors during physiological data acquisition from the database by hospitals. NLP will be used to find the common diseases over different geographical locations.*

Key Words: **Regional symptomatic disease; natural language processing; machine learning; physiological data acquisition; statistical data analysis.**

1. INTRODUCTION

Nowadays, healthcare is taken into account to be a significant challenge. Infectious diseases are among the most serious health issues in the world. The emergence of these diseases can be through air, water, direct contact with the infected person, biologically and ecological determinants [1]. In the year 2020, the world has witnessed an outbreak of infectious diseases, which is Corona Virus/ Covid-19. Around 64 L deaths globally are reported till date and it is getting multiplied at an awfully faster rate. Awareness of such infectious diseases needs to be spread widely among the people for the prevention of being infected in prior. These infectious diseases go on spreading over larger areas leading to epidemics and pandemics. Also, these outbreaks have major impacts on the population both socially and economically.

In the situation of epidemics/pandemics, when everything is virtual, there are many places in our country which lack medical facilities. The traditional way of treatment to disease may not be enough in the case of serious problems. Developing a medical diagnosis system based on Natural Language Processing (NLP) and Machine Learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis and preventing the spread for pandemics. Accurate and on-time analysis of any health related problem is vital for the prevention and treatment of the illness. Hence, detecting the spread of such epidemics / pandemics at an early stage across various locations is going to be helpful for early diagnosis and for death prevention on a bigger scale. After identification of an emerging pandemic, detecting the disease spread, local and international healthcare organizations may be notified earlier in order that they will take steps to halt the disease's progress [2]. Thus, controlling the epidemic diseases at the start of its spread may be a vital solution for epidemics/pandemics.

2. LITERATURE SURVEY

Harini D K, Natesh M [3], In this paper, machine learning algorithms is used for effective prediction of diseases. It uses both structured and unstructured data from hospital for effective prediction of diseases .Latent factor model is used to overcome the difficulty of missing data. A new convolutional neural network based multi-modal disease risk prediction (CNN-MDRP) algorithm is proposed in this paper. The proposed algorithm accuracy prediction reaches 94.8% than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

Shratik J. Mishra, Albar M. Vasi , Vinay S. Menon, Prof. K. Jayamalini [4] , The system implemented had the accuracy of 86.67% on the dataset of 120 patient data. The current system covered the general diseases or the more commonly occurring disease, so that early prediction and treatment could be done, and the fatality rate of deadly diseases decreases, with the economic benefit.

Minsung Kim, Joon Yeop Lee, Hwangnam Kim [5],This paper presents an Early Warning System (EWS) which is able to predict infectious disease outbreaks and detect the sudden increase of any livestock disease with the potentials to become epidemic before spreading.

Pahulpreet Singh Kohli, Shriya Arora [6], In this paper, different classification algorithms were applied, each with its own advantage on three separate databases of disease available in UCI repository for disease prediction.

3. PROPOSED WORK

The already existing system, for disease prediction, uses various data processing types, which is the actual basic foundation of Artificial Intelligence (AI) and Machine Learning (ML). Natural Language Processing (NLP) refers to one of the method of AI, which is concerned with giving computers a potential to understand text and spoken words in the same way humans can.

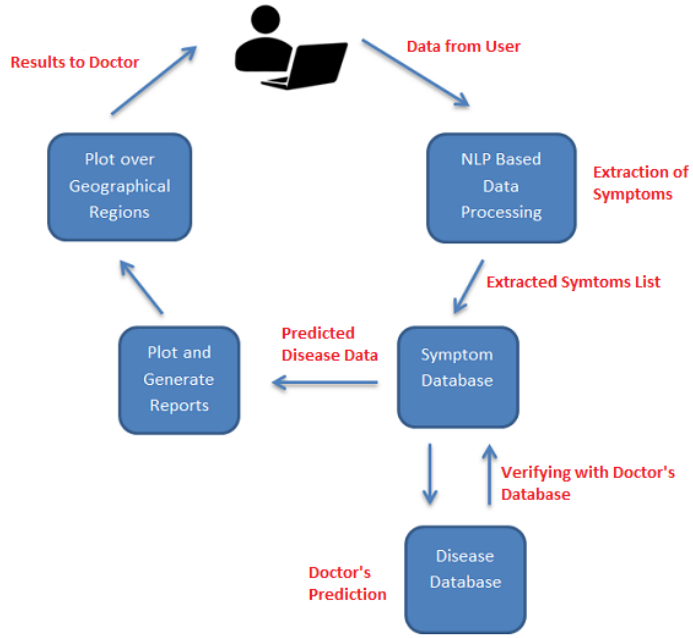

Figure 1: Proposed System Block Diagram

The Figure 1 shows the Block Diagram of the proposed System. It has mainly two parts. The first part is the 'disease prediction', that is using NLP algorithm to extract the symptoms from complaints given by user and then predicting the corresponding diseases. The second part is the 'plotting over geographical regions' which uses the latitude and longitude from different states, in order to know the spread of diseases.

The main aim of the proposed system is to develop an artificial intelligent system which detects the spread of diseases over geographical locations. This can be done by extracting the regional symptomatic data that is the symptoms, in the form of complaints, given by the paramedic/doctor during the physiological data acquisition. The database contains tables, which has the information about the patient details like name, age, gender, contact details, visit date, complaints/symptoms the patient having and location. Second table consisting of symptoms, disease lists and mapping of symptoms to disease prepared by the pre-registered doctors. Third table consists of city with latitude and longitude of location. The patient data will be stored on the web server so that the doctor can access the information whenever required from anywhere and does not have to be physically present.

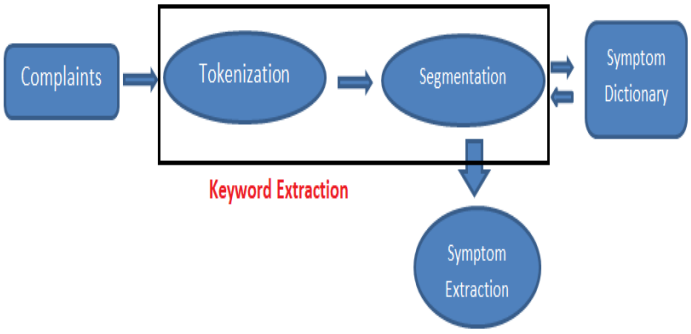

Figure 2: Symptom Extraction Process

The Figure 2 shows the symptom extraction process. So, the user enters the input in the form of complaints, which can be a sentence or a paragraph. Then the entered input will go through the tokenization process, where it breaks down the sentence or the paragraph into smaller chunks of words. This smaller chunks will be processed with 'stop word removal', to remove the stop words like 'is, this, the, are, a, an, etc.' After, the stop words are removed, the result is passed through removal of delimiters that is ' , and .' . Later when the delimiters as well as stop words are removed from the sentence/paragraph, the remaining words are referred to as 'candidate keywords'. Sometimes, the candidate keyword can also be the extracted symptom from the sentence/paragraph. The next is the segmentation process where meaningful phrases are created from the data obtained after the entire tokenization process.

And finally, the system will cross check the obtained segmented data and the symptom list in the database. Each time a match is found, will be considered as a symptom from the user's entered input. Later, these symptoms will be used for further disease prediction.

For the disease prediction, the database will have the list of symptoms and diseases which are prepared by the pre-registered doctors. After the process of symptom extraction, the system will make cross-matching with the database which contains both 'symptom and disease list'. Finally, the system will then predict the disease.

4. PLOTTING OVER GEOGRAPHICAL REGIONS AND GENERATING REPORTS

The final output is more beneficial when the system generates report in the form of graph. The graph indicates the rise of symptoms and disease location wise as well as month wise. The Figure 3 shows the month-wise symptom rise and the Figure 4 shows the state-wise symptom rise.

The system also generates rise of epidemics/pandemics over geographical regions. The geographical region on map uses

latitude and longitude from the database for mapping the symptoms over the states. After selecting multiple symptoms from the checkbox of the user interface, the map-view will show the number of patient spread across different regions having the selected symptom. Figure 5 shows the geographical view of the symptom "Cold and cough", "Skin problem" and "Body pain", in the states of India. This geographical view shows the total count of patients having particular symptom in the form of bubble map. Multiple symptoms on map can be distinguished by different color code. When the cursor moves over to the coordinates of any particular state or symptom, cursor will show the count of the patients in that state for that particular symptom, as displayed the count of patient for skin problem in the state Gujarat. The dataset used for geographical plotting consists of more than 7k entries.

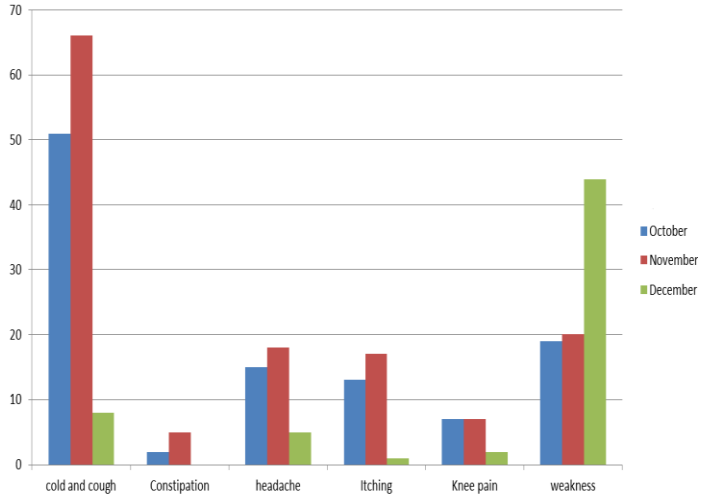

Figure 3: Month wise symptoms rise

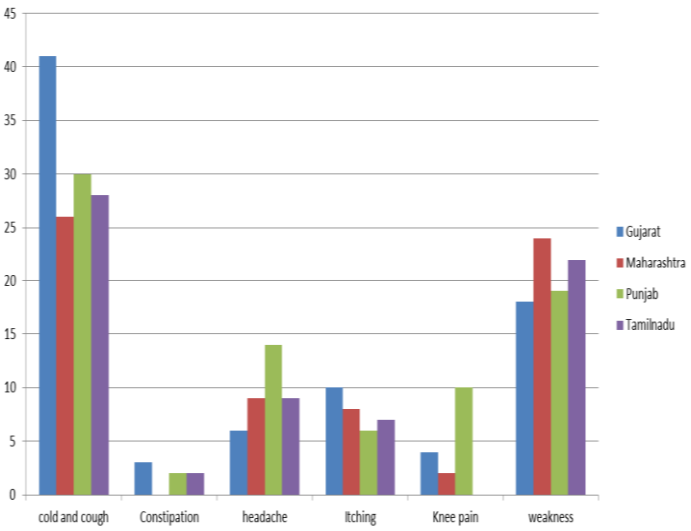

Figure 4: State wise Symptoms rise

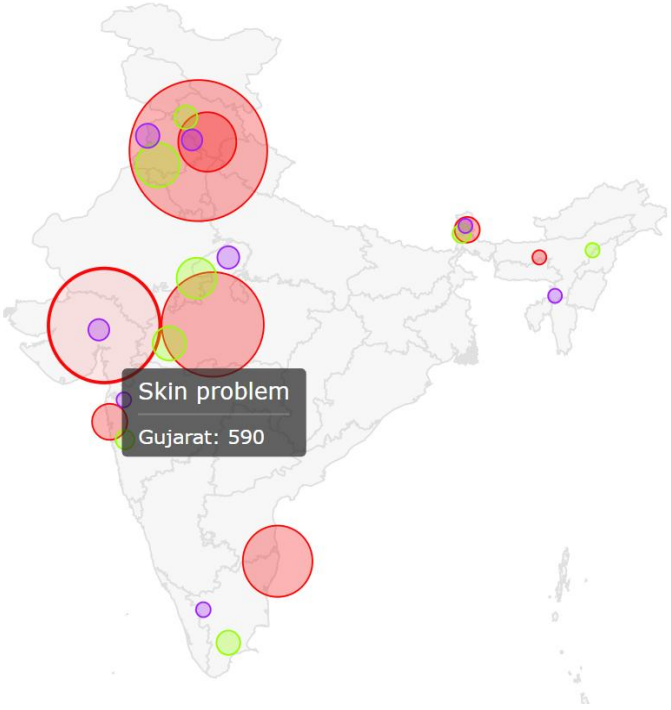

Figure 5: Symptom rise over geographical regions

5. CONCLUSION

The proposed system aims at detecting the spread of epidemics /pandemics over geographical map plot. The reports generated in the results shown can be used by medical or health organizations for analysis at national or international level. The future work is to visualize the data more prominently and add animations to the graphs. These graphs can consider many other attributes like year, age, gender, etc. Mapping over geographical regions can further be done at district level as well as village level.

6. REFERENCES

[1] Inayatulloh, Selvyna Theresia, "Early Warning System for Infectious Diseases", DOI: 10.1109/TSSA.2015.7440435 ,IEEE 2015.

[2] Khanita Duangchaemkarn, Varin Chaovatut, Phong tape Wiwatanadate, and Ekkarat Boon chieng, "Symptom-based Data Preprocessing for the Detection of Disease Outbreak", DOI:10.1109/EMBC.2017.8037393 ,pp. 2614-2617, EEE, 2017.

[3] Harini D K, Natesh M , "Prediction Of Probability Of Disease Based On Symptoms Using Machine Learning Algorithm" ,International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 05 | May-2018.

[4] Shratik J. Mishra, Albar M. Vasi , Vinay S. Menon, Prof. K. Jayamalini, "GDPS - General Disease Prediction System", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 03 | Mar-2018.

[5] Minsung Kim, Joon Yeop Lee, Hwangnam Kim, "Warning and Detection System for Epidemic Disease", DOI: 10.1109/ICTC.2016.7763517 ,pp. 478-483 ,IEEE, 2018.

[6] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction", 4th International Conference on Computing Communication and Automation (ICCCA), DOI:10.1109/CCAA.2018.8777449 ,pp. 1-4 , 2018.

[7] Fariz Bramasta Putra et. al., "Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems", International Electronics Symposium (IES), DOI: 10.1109/ELECSYM.2019.8901644, pp. 1-5 ,IEEE, 2019.

[8] Aswathy K P, Rathi R, Shyam Shankar E P, "NLP based Segmentation Protocol for Predicting Diseases and Finding Doctors", International Research Journal of Engineering and Technology (IRJET) , Volume: 06 Issue: 02 | Feb 2019.

[9] PM. Lavanya , E. Sasikala, "Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey", International Conference on Signal Processing and Communication, DOI: 10.1109/ ICSPC51351.2021.9451752, pp. 603-609,IEEE, 2021.