INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1977)
Available online at: https://www.ijariit.com

Machine Learning and NLP based System for Medical Data Analytics and Prediction

Pooja S. P.
poojasp.is18@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Harshitha H. N.
harshithahn.is18@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Meghashree M.
meghashreem.is18@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Navyashree A. M.
navyashreeam.is17@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Merin Meleet
merinmeleet@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

ABSTRACT

In the present technical era, healthcare providers generate large amounts of medical data on a day-to-day basis. Produced clinical information is placed away carefully as Electronic Health Record (EHR) as essential information archive of medical clinics. These days Artificial Intelligence (AI) have been developing rapidly in recent years. Particularly, wellbeing data framework can get most advantages from the AI benefits. Specifically, symptom based disease prediction expectation exploration and creation turned out to be progressively mainstream in the medical care area as of late. In the paper, we have proposed a structure to assess the proficiency of applying both Natural Language Processing (NLP) and Machine learning (ML) advances for disease prediction framework. As an example we have interpreted n2c2 heart related disease symptom datasets from DBMI portal. The acquired patient records is in XML format which is parsed and converted to structured format and naive Bayes algorithm is applied for training.

Keywords: *Machine Learning, Artificial Intelligence, Natural Language Processing, Data Extraction, EHR, Naive Bayes*

1. INTRODUCTION

Medical care data sets are developing dramatically, and Natural language processing system convert this data into reverence. Clinical services providers, drug organizations and biotechnology firms all utilization Natural Language Processing to develop patient outcomes, Medical text contains info about diagnoses, treatments, symptoms, drug use for patient, can be utilized to progress medical care for further patients. The doctor also writes patient perceptive for the inference of the conclusion of patient in patient record. These days, Machine learning algorithms have gotten vital in the clinical area, particularly for diagnosing infection from the clinical data set. Numerous organizations utilizing these procedures for the early forecast of infections and improve clinical diagnostics. AI based applications can deliver useful information or can also analyze early based on simple datasets. For example symptom and relational datasets.

Previous analysis like, by using Naive Bayes classifier, predicting conditions like skin infections and predicting presence of swine flu, prediction of various cancer related issues using clustering and machine learning techniques have also been carried out and all these predictions have given considerably quality results. Although good results have been produced almost all the analysis are reduced to some diseases or certain clinical conditions which are important in medical framework. However, few analysis have come up with new methods from which medications can be predicted by given symptoms, but accuracy and reliability are yet issue to be discussed. The research that we have done focuses on unstructured data i.e., n2c2 dataset which is in XML format is converted into structured format. The Machine Learning algorithms are applied to structured data and based on the symptoms, the medication is provided. This aims to help doctors and practitioners in predicting the medication for the given symptoms using Machine Learning techniques. Our research has achieved 70% accuracy.

2. LITERATURE SURVEY

Hong Qing Yu [3] has conducted research to create an efficient disease prediction model by applying new methods as tentative case study to inspect the excellence and problems of the results in direction to achieve upcoming research ways. The suggested architecture will analyze, investigate current solutions and build efficient model which can be used to predict infections and medications based on the provided symptoms. Author uses NLP, K-Means and Naive Bayes Algorithms to analyze 298 medical records and have accuracy around 80% and proposed model can

reach up to 93%.

Shetty, Karthik, & Ashwin [5] has made use of most normally exhibited disease datasets, Author has develop a technique to predict the related infection formed on the input given in form of manifestations. The presented framework makes use of the abilities of various ML algorithms incorporated with script processing to attain precise predictions. In this research author has made use of the datasets which are assessable online to carry through the predictions on test data via ML. The goal of the work is to develop a model and a functional design for prediction ODF infections considering different manifestations. For this plan, the datasets which contain details regarding 200 most common infections and the symptoms which are associated with the infections are considered. The datasets used has gone under a lot of preprocessing. The initial stage was to covert the data set into a table with symptoms as the fake variables and infections as the mark column on which systems are trained. The datasets has been tested with 3 algorithms Naive Bayes, Random Forest, Decision Tree, Pandas data frames and scikit-learn library of python programming Language has been used to implement algorithms and also to process the data. This experiment achieved around 99% accuracy.

Pattekari and Parveen [4] accompanied a research to forecast heart disease using data mining techniques and Machine Learning, the purpose of this experiment is to build a smart support framework for the physicians which can predict heart related diseases using Naive Bayes Algorithm, based on the given set of input features. Author has deduced by claiming that the planned framework is most precise solution to heart related issues in patients. However, author has not used any real-words datasets or any real words use cases to support the statement and also there is no proof to support the report. Thus, the question arises is the reliability.

Authors Kunjir, Sawant and Shaikh [11] highlights the application of classifying and predicting disease by implementing the operations on medical data generated in the field of medical and healthcare. This project has implemented naïve Bayes algorithm to predict particular disease by training it on certain data set. J48 and Naive Bayes algorithms are used for comparing accuracy testing and performance, also for amount of time that is consumed for training the medical datasets.

Tikotikar & Kodabagi[12] has conducted a survey and discussed about the decision constraint, and features used for calculating the disease. Most of the datasets considered in so many existing techniques are associated to breast cancer and heart. The paper also focuses on significance of various classification approaches for disease prediction in clinical datasets. The different data mining methods are used as classifier, to develop a effective system for infection prediction. It is agreed by the through survey that acquiring the essential information from the clinical data helps to maintenance well cognizant diagnosis and conclusions.

Chodey and Gongzhu Hu [17] has made use of supervised ML structure for estimate of named entities based on the conditional random field design. Task7 organizer datasets are used to assess the framework. In this paper, author has discussed regarding the tool that is built for NER that brings out disorder reference after evaluating huge quantity of medical data and mapping which reference of conditions to Unified Medical Language System CUIs. The goal of Semantic Evaluation Tasks is to evaluate the execution of semantic analysis with the help of tools. Among the different Tasks defined by the Semantic Evaluation, author has selected Task7 to develop and assess the execution of the tool. The framework was applied on the datasets of medical records which were issued by the Semantic Evaluation organizers. The outcome shows that the framework produce less F-score and precision in strict setting than in the relaxed setting. The developed structure is to bring out different attributes for recognition of ideas in Unified Medical Language System. A CRF system is developed on the training datasets and assessed on the testing datasets. The attributes to bring out the data is described. The datasets include 2 sets that is 200 files of training data and 100 files of testing data are divided into four subsets: Echo, ECG, Radiology and Discharge summary.

3. METHODOLOGY

In this paper, an architecture has been proposed for the execution of a method that predicts medications based on symptoms. Here patient records are considered as unstructured data which we have collected from DBMI data portal and the dataset is n2b2. In n2b2 we have collected heart related records. The most of these Medical NLP datasets were initially created at a earlier NIH-supported National Center for Biomedical Computing (NCBC) recognized as i2b2: These datasets now persist below the supervision of the Branch of Biomedical Informatics at Harvard Medical School.

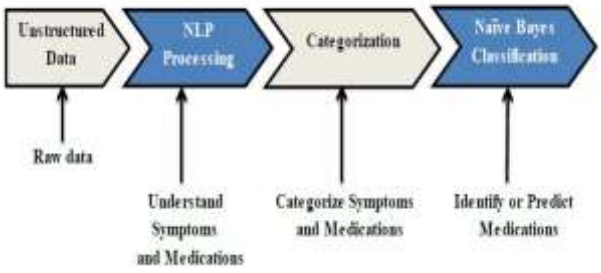

Fig.1: Flow Diagram

The dataset is in XML format, the first step is analyzing the XML file. XML analyzing is the procedure of understanding an XML file and furnishing an affiliate to the user presentation for retrieving the file. Python uses XML module for parsing the XML files and the main focus will be on the ElementTree XML API of this module. By reviewing the raw files, we detected that there is huge amount of frequent situation names with dissimilar symptoms. Consequently, we want to parse the raw files to support Natural Language Processing, NLTK is a module used for developing programs that effort with human language datasets for relating in natural language processing (NLP). The Natural Language Processing works are developed by programming with Python's natural language toolkit library. The procedure includes four approaches 1.) Tokenization 2.) Removing Stop Words 3.) POS tagging 4.) Chunks.

Tokenization is a direct process where paragraphs are converted to sentences and sentences to words. In the datasets the symptoms and medication for every state are present in sections. Using the Natural Language Toolkit tokenization approach, the sections are split into words of sentences creating it to use for the algorithm.

Stop word deletion is a process to remove the unwanted tokens in data. Example, stop words like ‘to’ and ‘of’ contains no sense and not significant in medical related projects. Adding to this, filler word deletion will decrease the proportional space and deliver benefit by collection of tokens.

Parts of speech tagging refers to allocating parts of speech to each words in a sentence, unlike phrase matching, which is fulfill at the sentence or multi-word level, parts of speech tagging is executed at the token level. Medications and symptoms can be easily identified whether they are clinical words or not. For example if the Input = "seen in cardiac" then output=[(seen=NNP), (in=IN), (cardiac=NNP)], NNP refers to proper noun, IN refers to preposition. By this it is easy to group symptoms and medications.

Chunking is the activity of grouping akin words together based on the existence of the word. In our implementation grammar is considered by the chunk need be produced. The grammar suggests the arrangement of the expression like nouns and pronouns etc. These will be go with generating the chunks. By considering above example, here NNP are grouped together which will help us in further processing.

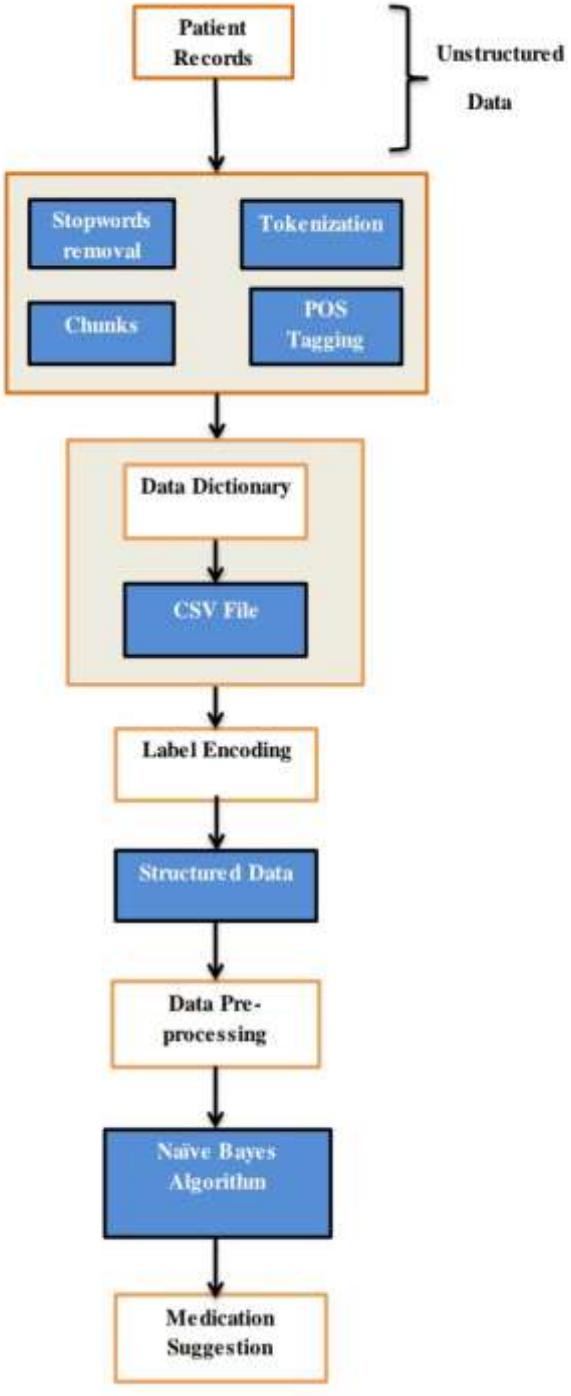

Fig. 2: System Architecture

It is necessary to extract clinical terms from patient records and categorize them into symptoms and medications. To execute this we manually studied all the patient records from the datasets and generated a data dictionary which categorizes symptoms and medications. After performing all these procedures the symptoms and medications are imported to the CSV file.

Next step is Label encoding the CSV file. Since the CSV file we have found has values in string format, it is necessary to label encode. Label encoding in Python, we replace the definite value with number value among 0 and the sum of classes -1. If the symptoms and medications are present in the CSV file, it is encoded with 1 else 0. At present the data we have is in structured format and can be used for pre-processing and training.

While pre-processing we have processed the data by removing null values and the structured data is given ML algorithm for training.

And the final step is training a structured data by using ML algorithm, we have used naïve Bayes classification algorithm for training the data. Naive Bayes is a general ML algorithm mainly used for script classification. It makes use of selective classifiers created on the values of Bayes Theorem. It executes based on the variables in a records. Though giving new features, the system will calculate the possibility of the comeback variable datasets belonging to a certain class. After training the input is given in the binary format, medications are predicted based on the given input. And we have achieved the accuracy of 70%.

4. RESULTS AND DISCUSSIONS

A confusion matrix is a table that can be used to evaluate the performance of machine learning algorithm, basically a supervised learning. Each column of the confusion matrix signifies the instances of a predicted class and each row signifies the instances of a actual class.

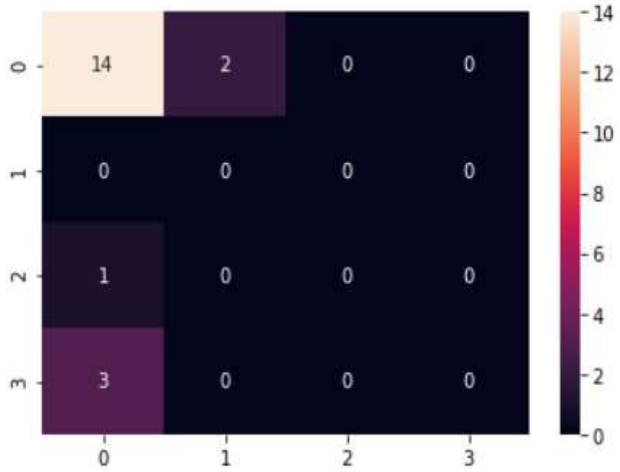

Fig. 3: Heat map

There are two possible predicted classes: yes and no. In our case if the medication is correctly predicted for the symptom it is yes else it is no. True Positives (TP): cases in which the prediction is yes. True Negatives (TN): cases in which the prediction is no. False Positives (FP) or Type I error: Prediction was yes but the result was no. False negatives (FN) or Type II error: Prediction was no but the result was yes.

	Precision	Recall	f1-score	support
0	0.78	0.88	0.82	16
2	0	0	0	0
3	0	0	0	1
4	0	0	0	3
accuracy			0.7	20
Macro avg	0.19	0.22	0.21	20
Weighted avg	0.62	0.7	0.66	20

Fig. 4: Classification report

4. CONCLUSION AND FUTURE ENHANCEMENTS

In the paper, we have introduced an tentative evaluation study system for building a symptoms based system for medical data and analysis. We took a 200 files from n2c2 data set from the DBMI portal and applied Natural Language Processing techniques and Machine Learning algorithms in the system. The accuracy achieved is 70%. Still, there are several issues about this study area, some are lack of knowledge, improper unstructured datasets and segmentation problems. Therefore, it is necessary to monitor our study to guidelines of acquiring proper datasets and categorizing the data.

There are many advancements which can be done on this project some are, 1. This system can be used to predict other diseases, clustering method can be used to categorize other diseases and the complete module can be used to predict medications for most of the diseases. 2. Online portal chat box can be developed when a user provides their symptoms there are two outcomes where user gets either medications for given symptoms or will be advised to consult a doctor. 3. The current system design can be enhanced to be used as a mobile application, So that doctors can assist patients.

5. REFERENCES

[1] S. M. Shah and R. A. Khan, "Secondary Use of Electronic Health Record: Opportunities and Challenges," in IEEE Access, vol. 8, pp. 136947-136965, 2020, doi: 10.1109/ACCESS.2020.3011099.

[2] S. M and A. Chacko, "A Case for Semantic Annotation Of EHR," *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, Madrid, Spain, 2020, pp. 1363-1367, doi: 10.1109/COMPSAC48688.2020.00-66.

[3] H. Q. Yu, "Experimental Disease Prediction Research on Combining Natural Language Processing and Machine Learning" *2019 IEEE 7th International Conference on Computer science and Network Technology(ICCSNT)*, Dalin, doi:10.1109/ICCSNT47585.2019.8962507.

[4] S. Pattekari and A. Parveen, (2019). "Prediction System for Heart Disease Using Naïve Bayes". International Journal of Advanced Computer and Mathematical Sciences, 3(3),pp.290-294.Available at: https://pdfs.semanticscholar.org/ d32e/ e90a5de89093a4fc95f43e0409cb91414726. pdf [Accessed 31 May 2019]..

[5] S. Vijava Shetty, G. A. Karthik and M. Ashwin, "Symptom Based Health Prediction using Data Mining," *2019 International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2019, pp. 744-749, doi: 10.1109/ICCES45898.2019.9002132.

[6] M. R. Mia, S. Akhter Hossain, A. C. Chhoton and N. Ranjan Chakraborty, "A Comprehensive Study of Data Mining Techniques in Health-care, Medical, and Bioinformatics," *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, 2018, pp. 1-4, doi:IC4ME2.2018.8465626.

[7] S. Kawthankar, R. Joshi, E. Ansari and S. D'Monte, "Smart Analytics And Predictions For Indian Medicare," *2018 International Conference on Smart City and Emerging Technology (ICSCET*), Mumbai, India, 2018, pp. 1-5, doi: 10.1109/ICSCET.2018.8537383.

[8] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," *2018 24th International Conference on Automation and Computing (ICAC)*, Newcastle Upon Tyne, UK, 2018, pp. 1-6, doi: IConAC.2018.8748992.

[9] R. Maheshwari, K. Moudgil, H. Parekh and R. Sawant, "A Machine Learning Based Medical Data Analytics and Visualization Research Platform," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-5, doi: 10.1109/ICCTCT.2018.8550953.

[10] N. E. A. Amrani, M. Youssfi and O. E. K. Abra, "Semantic interoperability between heterogeneous multi-agent systems based on deep learning", *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 1-6, 2018

[11] A. Kunjir, H. Sawant and N. F. Shaikh, "Data mining and visualization for prediction of multiple diseases in healthcare," *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, Chirala, Andhra Pradesh, India, 2017, pp. 329-334, doi: 10.1109/ICBDACI.2017.8070858.

[12] A. Tikotikar and M. Kodabagi, "A survey on technique for prediction of disease in medical data," 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), 2017, pp. 550-555, doi: 10.1109/SmartTechCon.2017.8358432.

