International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

NLP BASED CLINICAL DATA ANALYSIS FOR ASSESSING READMISSION OF PATIENTS WITH COPD

Priyanka V. Medhe[1], Dinesh D. Puri[2],

[1]Student, Department of Computer Science and Engineering, SSBT's College of Engineering and Technology, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, M.S, India
[2]Assistant Professor, Department of Computer Science and Engineering, SSBT's College of Engineering and Technology, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, M.S, India

---

Abstract - *Predicting the readmission of patients from medical text has been a field of ongoing research. The main motivation for this project is the availability of enormous amount of data that could effectively help in medical research. Processing of these data will provide information that would aid in the research subject to readmission process. Though research to predict hospital readmission has begun to appear in the last two years, no research has explored Chronic Obstructive Pulmonary Disease (COPD) patient readmission, regardless of methodology. The proposed system employs regular expression based approach to predict COPD patient readmission. The system works in two phases, training and testing. Generation of regular expression is done in training phase. In testing phase, regular expressions are generated of the test data and then matched with trained data for prediction. The Experimental Results shows the accuracy of the prediction of readmission of COPD patients, by calculating the Precision, Recall and F-Measure of the proposed system.*

Key Words: **Chronic Obstructive Pulmonary Disease, Natural Language Processing, Readmissions, Regular Expressions, Prediction.**

1. INTRODUCTION

The concept of natural language processing is to develop a computer system that can analyze, understand and synthesize natural human languages. Under the domain of artificial intelligence natural language falls with the goal of understanding and creating meaningful expressions in the human language. Natural Language Processing (NLP) involves both computer understanding and computer generation of Natural Language text. The goal is to enable Natural Languages such as English, French or Japanese, etc. which serves as the medium of interaction between the users and computer systems (Natural Language Interaction) or as the object that a system processes into automatic text translation or summarization (Natural Language Text Processing). In the analysis of Natural Language (NL), the task is to translate an utterance, often in context, into a formal specification that the system can process further. In NL Interaction, such further processing may involve factual data retrieval and/or reasoning, as well as generation of an appropriate response. In Text Processing, generation of an appropriate translation or a summary of the original text(s) is done on the basis of understanding the text first, or the formal specification may be stored for later, to serve as the basis for more accurate document retrieval. In NL Generation, the task is to convert some representation of what the system wants to communicate (e.g., a fact, an explanation, a description or definition) into a clear, well-structured, and hence understandable piece of NL prose. A readmission is the next subsequent admission of a patient as an acute admission within a defined period of time. Readmissions would consume resources for a given population in a time frame, increasing the cost for overall medical treatment.

1.1 Overview of Natural Language Processing

Natural language processing (NLP) approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. 1. Symbolic Approach: Symbolic approaches are based on human developed rules and lexicons. The basis behind this approach is the representation of facts about language through well understood knowledge representation schemes and associated algorithms. They perform deep analysis of linguistic phenomena. 2. Statistical Approach: Statistical approaches are based on observable and recurring examples of linguistic phenomena. This approach uses various mathematical techniques to analyze recurring themes large text corpora. Machine learning based NLP solutions use this approach. 3. Connectionist Approach: Connectionist approach develops generalized models from examples of linguistic phenomena. However, connectionist models combine statistical learning with various theories of representation thus the connectionist representations allow transformation, inference, and manipulation of logic formulae.

Predicting the readmission of patients from medical text has been a field of ongoing research. The main motivation for this project is the availability of enormous amount of data that could effectively help in medical research. These data are mostly available as free text collected through research applications. Processing of these data will provide information that would aid in the research subject to readmission process. This could be achieved by filtering the criteria from the free text to be used in the database queries. Direct comparison of methodologies in the field of health informatics can be difficult. Unlike other domains, data is often restricted and cannot be released publicly. Since the data is unavailable, existing system uses a mix of bag-of-words methodology and extension of Clinical Text Analysis

and Knowledge Extraction System (cTAKES) annotations using feature selection methods. The existing system requires complete observations, if some input variables are missing then it is not possible to predict the readmission of patients. Due this ambiguity and complexity, an architecture which can devise the readmission prediction of patients by applying regular expression based approach is required.

Contribution in this work is the training and testing of the data by utilizing Regular Expression (RE) based approach, matching the regular expressions generated of the test data with the trained data, and then predicting the readmission of COPD patients by considering the matched data.

The rest of the paper is organized as follows: Section II describes Related Work. Section III describes Methodology. Section IV describes Result and Discussion, while Section V describes Conclusion and Future Work of the paper.

2. RELATED WORK

Weng et al., in [1], have developed a processing framework for eligibility criteria extraction and representation (Elixr). The Elixr system consists of semantic pattern mining and syntactic tree parsing to generate semi-structured eligibility criteria.

Doing-Harris et al., in [2], had used the clustering algorithm with semantic types and vocabulary for their data representation, to perform the unsupervised learning task across different note types and different document sources, and yielded good performance for identifying clinical sub languages.

Hughes et al., in [3], proposed a framework in which convolutional neural networks (CNN) is applied with distributed word representation to medical text classification task at a sentence-level and yielded competitive performance. CNN or a variant of recurrent neural network or Long Short-Term Memory (LSTM) has also been applied at the document-level to learn semantic representations in documents for general sentiment analysis. At the character-level CNN has also been applied for different text classification tasks.

Dr. Breiman et al., in [4], proposed the random forest algorithm in 2001, which has been extremely successful as a general purpose classification and regression method. Excellent performance in settings has shown by the approach, which combines several randomized decision trees and aggregates their predictions by averaging, where the number of variables is much larger than the number of observations.

Milian et al., in [5], have developed an approach to the formalization of free-text eligibility criteria using pattern detection. They created 165 patterns that cover conditions related to demographic information, disease characteristics and prior and concurrent therapies. Their pattern detection algorithm uses regular expressions to identify specific patterns in clinical trial eligibility criteria.

Duggal et al., in [6], uses Apache cTAKES to annotate the unstructured EHR. This research specifically looks at the 30 – day readmission rate of diabetes patients in an Indian hospital. The data contains 0.129 readmission rate and 9,381 instances. Several machine learning algorithms were compared. The highest AUC for this diabetes study was 0.688 using Random Forests. The results are typical of readmission analysis.

Hao et al., in [7], have created an algorithm that uses heuristic rule and pattern learning to extract and normalize temporal expression in free-text eligibility criteria. All these work were done with the help of eligibility criteria from ClinicalTrials.gov.

3. METHODOLOGY

The proposed system focuses on the Readmission of patients. Identifying the readmission of patients is a difficult task, because evaluating the attributes of the underlying content and understanding the varying ranges of symptoms are more critical. Readmission prediction is strongly connected with different symptoms which the patients are showing. While executing the project, dataset containing the patient symptoms as attributes is used. The proposed approach uses regular expression based approach in order to facilitate and improve rule based model to predict the readmission of COPD patients. The model uses medical dataset that is available as a central resource for matching the testing dataset with the trained dataset. Preprocessing is applied to the training dataset. Then Regular expressions are generated of this preprocessed training dataset. Testing dataset is used of which regular expression generation is done. After this regular expression matching is done on the generated regular expressions of testing dataset.

3.1 Architecture

Figure 1 show the architecture of the proposed system, which consists of two phases: training phase and testing phase.

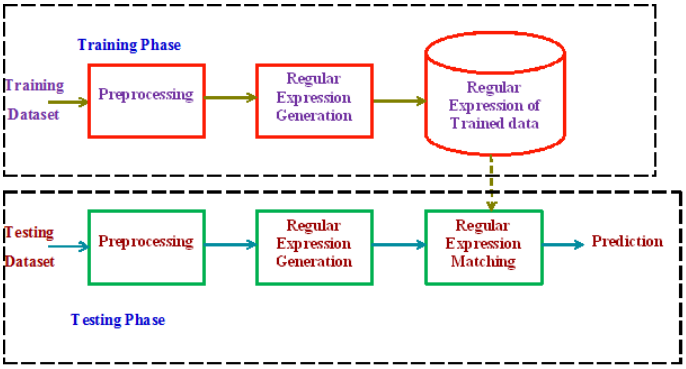

Fig -1: Architecture of the Proposed System

In training phase dataset is taken for preprocessing. In preprocessing the symptoms with COPD patients are only chosen. After that the regular expressions are generated of the preprocessed data. These regular expressions which are generated are the trained data. Later in a testing phase another dataset is taken for testing data. Then generation of

regular expression is done on the testing dataset. Later on these generated regular expressions are matched with the regular expression of trained dataset. And from this it is predicted whether the readmission is required or not for the patient.

3.2 Preprocessing

Dataset is taken as input for preprocessing. This data preprocessing is done in the training and testing phase of the process. In preprocessing the attributes are generated in the form of 0's and 1's, by setting a threshold value to them. Here a particular threshold value is set for all the data ranges that are given in attribute form.

3.3 Regular Expression Generation

This regular expression generation is done both in training phase and testing phase. The input required for regular expression generation is taken from preprocessing. The input given to the regular expression generation is in the form of array of attributes that does not contain any raw data. A regular expression generation algorithm is applied to the preprocessed data and then the regular expressions are generated of the preprocessed data.

By applying certain conditions on records, regular expressions are generated. For example, records 10101011, 00101001 and 10110100 follow certain conditions then the regular expressions generated of this records are 10+101?01*, 0+1010*1?, 101+010*. In this way each record in the dataset which follow the conditions generates regular expression. This process is applied to both training dataset and testing datasets. In testing dataset the only difference is to select don't care attributes. Then after this selection the regular expressions are generated of each record of testing dataset. Here the symbols "*, +" acts as repeaters.

_ Plus symbol (+): It represents repeatation of the preceding character (or set of characters) for atleast one or more times (upto infinite).

_ Asterisk symbol (*): This symbol matches the preceding character (or set of characters) for 0 or more times (upto infinite).

_ Optional character (?): This symbol represents that the preceding character may or may not be present in the string to be matched.

The Algorithm 1 focuses on the generation of regular expressions of training dataset. It converts the attributes of the input dataset into required form. A result string is created of this obtained attribute values. This string is stored in the form of char array. All char array are taken for processing. Then it checks whether the char is repeated or not. If char is repeated then how many times it is repeated that number is added to the output string, if there is already a number showing repeatations in the string then next char is taken into consideration and if char is not repeated then this char is added in output string. This generates the output string. And this strings generated are the trained data.

Algorithm 1: Regular Expression Generation Algorithm for Training Data:

1: Procedure Preprocessing.
Require: Training Dataset containing symptoms of patients.
2: Input: Dataset that contains the number of symptoms in the required form.
3: Output: String of binary numbers.
4: While (readline! = EOF).
5: Parse the attributes based on conditions.
6: Generate results for all attributes.
7: Create result string for attribute values.
8: End While.
9: For all records r.
10: Split record in single characters c.
11: For all characters c.
12: Check if the character c is repeated.
13: Increment number of time repeated.
14: Check if character c is new and previous character is repeated.
15: Print number of times repeated.
16: Combine complete string.
17: End For.
18: End For.
19: Store generated regular expression.

The Algorithm 2 focuses on the generation of regular expressions of testing dataset. It converts the attributes of the input dataset into required form. A result string is created of this obtained attribute values. Then Don't Care attributes are selected from this result string. The same process of generation of regular expression is performed on the string created after the selection of Don't Care attributes. And then the regular expressions are generated of this testing data. Further these generated regular expressions are carried for matching with the trained data.

Algorithm 2: Regular Expression Generation Algorithm for Testing Data:
1: Procedure Preprocessing.
Require: Testing Dataset containing symptoms of patients.
2: Input: Dataset that contains the number of symptoms in the required form.
3: Output: String of binary numbers.
4: While (readline! = EOF).
5: Parse the attributes based on conditions.
6: Generate results for all attributes.
7: Select Don't Care attributes.
8: Create result string of these attribute values.
9: End While.
10: For all records r.
11: Split record in single characters c.
12: For all characters c.
13: Check if the character c is repeated.
14: Increment number of time repeated..
15: Check if character c is new and previous character is repeated.
16: Print number of times repeated.
17: Combine complete string.
18: End For.

19: End For.
20: Store generated regular expression.

3.4 Regular Expression Matching

Regular expression matching is done only in testing phase to predict the readmission of patients. In testing phase also regular expressions are generated of the input data. Generated regular expressions are in the form of 0 and 1. If the generated expression of the testing data is matched with the trained data then we predict that readmission is required. The given string is matched when there is no difference between the regular expression generated of the testing data and regular expression generated of the trained data.

For example, the regular expression of the trained data is 10+101?01*, 0+1010*1?, 101+010* and the regular expression of testing data is 0+1010*1?, 101+010*, 10*1?0+1, 0*1010?1+ then it shows that out of four records 2 records are matched and 2 records are not matched. From this matching prediction of readmission is made. When no don't care condition is applied to the testing data and if the generated regular expressions are fully matching with the trained data, then it is that the record is matching 100 percent and the patient requires readmission.

If we apply the don't care conditions on the testing data, means we consider some attribute as don't care, then the matching shows that how many percent the regular expression generated is matched with the trained data. If the expressions generated are matched above the expected value then it is predicted that the patient requires readmission.

For example, the regular expression of the trained data is 10+101?01* and the regular expression of the record 1001010010 of the testing data is 1010+1 by considering the 3rd,6th and 10th attribute as don't care, it shows that the tested regular expression is matching above expected value. It means that the patient having this record requires readmissions.

The Algorithm 3 shows how the regular expression matching of the testing data is done with the trained data. This regular expression matching with the trained data makes the prediction for readmission of patients easy.

Algorithm 3: Regular Expression Matching Algorithm:
1: Procedure Matching.
Require: Read regular expression generated from training phase.
2: Input: Generated regular expressions of the test data.
3: Output: String of binary numbers.
4: If expression is blank then
continue
5: initialize max as 0
6: If first regular expression R1 length is less than second regular expression R2 then
max = length(R2)
7: Else
max = Length(R1)
8: matches R1 and R2
9: For i 1 to 10
10: If R1[i]=R2[i]
true++
11: Else
false++
12: End For
13: Percentage = true/ max
14: If percentage is greater than 90
Matched
15: Else
Not matched

4. RESULT AND DISCUSSION

Performance Metrics are used for predicting the readmission of COPD patients. Datasets containing records of COPD patients are used for readmission prediction. Each record has score which is calculated through following Equations. The Precision, Recall and F-measure are used for readmission prediction.

Precision is defined as the ratio of the number of relevant records retrieved to the addition of the number of relevant records retrieved and number of irrelevant records retrieved. Calculating the Precision is a measure of number of correct records penalized by number of incorrect records. It is a fraction of retrieved correct records that are relevant to find. For precision, the formula given in below Equation is used.

$$\text{Precision} = \frac{\text{Relevant Retrieved Records}}{\text{Total Retrieved Records}}$$

Recall is defined as the ratio of the number of relevant records retrieved to the addition of the number of relevant records retrieved and number of relevant records not retrieved. Calculating the Recall is the measure of number of correct records as number of missed entries. Recall is the measure of the ability of a prediction model to select instances of a certain class from a dataset. It is also called sensitivity, and corresponds to the true positive rate. For recall, the formula given in below Equation is used.

$$\text{Recall} = \frac{\text{Relevant Retrieved Records}}{\text{Total Relevant Records}}$$

F- Measure is a measure of testing accuracy. For the F-measure calculation, the precision and their respective recall are considered. By applying the formula given in below Equation F-measure values are calculated. It considers both the precision and recall of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. F-measure is the weighted average of the Precision and Recall. The F-measure or the Balanced F-measure is the harmonic mean of Precision and Recall.

$$\text{F-Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Experimental results consists the values of Precision, Recall and F-measure with respect to proposed system, based on the datasets containing 100, 200, 300, 400, 500, 600 records of patients as input dataset.

The experimentation results shows the number of total retrieved records from the datasets, relevant retrieved records of the datasets and the total relevant records of the datasets. The value of this experimentation is shown in Table 1.

Table -1: Experimentation Results

No. of Records	100	200	300	400	500	600
Total Retrieved Records	88	180	272	361	449	536
Relevant Retrieved Records	83	172	261	348	435	252
Total Relevant Records	92	190	285	370	462	555

For calculating Precision and Recall the formulas given in the precision and recall equations are used. These equations are applied to all records given in the dataset and then precision and recall of them is calculated. Chart 1 is plotted by considering precision of all dataset and chart 2 is plotted by considering recall of all dataset. For calculating the F-measure throughout, precision and recall of all records are used from dataset. Chart 3 is plotted by considering the F-measure of all datasets.

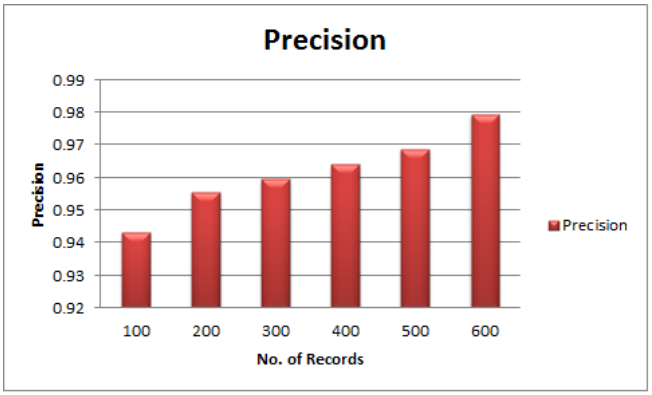

Chart -1: Precision

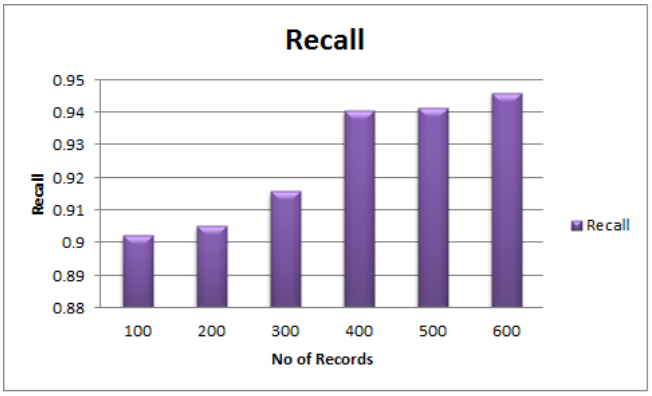

Chart -2: Recall

Table 2 shows the precision, recall and F-measure for the prediction of readmission of patients of the proposed system.

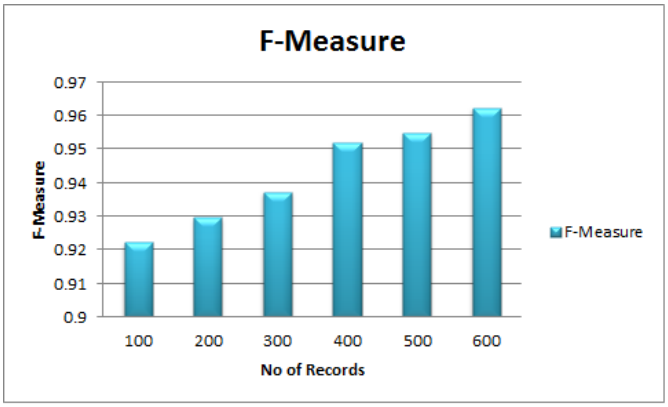

Chart -3: F-measure

Table -2: Precision, Recall and F-measure

No. of Records	100	200	300	400	500	600
Precision	0.94318	0.95556	0.95956	0.96399	0.96882	0.97948
Recall	0.90217	0.90526	0.91579	0.94054	0.94156	0.94595
F-measure	0.92222	0.92973	0.93716	0.95212	0.95499	0.96242

The number of relevant records retrieved, number of total relevant records and the number of total retrieved records are observed. Further by applying the formulas given in above equations, the Precision for readmission prediction is like relevant retrieval/total retrieval, where the relevant retrieval is 83 for the dataset containing 100 records and total retrieval is 88 from which the precision of 0.94318 is calculated. Also Recall is calculated like relevant retrieval/total relevant, where the relevant retrieval is 83 and total relevant is 92 from which the recall 0.90217 is calculated. Each calculation uses the same procedure for all datasets. At last through the values of respective precisions and recalls, the F-measure is calculated. The F-measure of dataset containing 100 records is 0.9222. It is calculated by considering it's precision and recall. Result of experiment states that the rate of accuracy of proposed system is more precise. The precision graph gradually increases as the number of records increases in the dataset. Similarly the recall graph increases as the number of records increases in the dataset. F-measure also gets changed according to the precision and recall. The F-measure graph also increases as the number of records increases in the dataset. The F-measure values are calculated from the precision and recall measures. The experimental result shows that readmission of COPD patients is predicted in a more accurate manner by applying regular expression based approach. The precision and recall gets improved as the number of records in the

dataset gets increased. As a result of increase in precision and recall F-measure also gets increased.

5. CONCLUSION AND FUTURE WORK

Prediction of readmission of patients is a challenging task for the regular expression based approach. The system is able to predict patient readmissions using regular expression generation and matching approach, which is better than existing systems. This approach offers the advantage that separate data collection is not required for readmission prediction since a large number of data is available by medical institutions. Training of structured models using regular expression generation approach showed potential feasibility for automatically increasing the size of training data in clinical text classification tasks. Testing data by matching the generated regular expression with it showed the better prediction policy. The experimental result shows that how accurately the readmission of COPD patients is predicted, by calculating the Precision, Recall and F-Measure of the proposed system.

In future, the proposed system may intents to extend efforts to predict the readmission of patients having multiple diseases as records become available.

ACKNOWLEDGEMENT

I express my deep sense of gratitude to the Prof. Dr. K. S. Wani (Principal of SSBT's College of Engineering and Technology), and Prof. Dr. Girish K. Patnaik (HOD of Department of Computer Engineering) and Prof. Mr. Dinesh D. Puri for their cooperation and guidance in completion of this work.

REFERENCES

[1] C. Weng, X. Wu, Z. Luo, M. R. Boland, D. Theodoratos, and S. B. Johnson, "Elixr: an approach to eligibility criteria extraction and representation." Journal of the American Medical Informatics Association, 18(Supplement 1), pp. 116-124, 2011.

[2] K. Doing-Harris, O. Patterson, and S. Igo, "Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts." In: Proceedings of the 7th international workshop on Data and text mining in biomedical informatics, pp. 9-12, November 2013.

[3] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks." Studies in Health Technology and Informatics, pp. 235-246, 2017.

[4] L.Breiman, "Random forest," Machine Learning, pp. 5-32, 2001.

[5] Milian, Krystyna, A. Bucur, and A. T. Teije, "Formalization of clinical trial eligibility criteria: Evaluation of a pattern-based approach." Bioinformatics and Biomedicine (BIBM), IEEE International Conference, pp. 1-4, 2012.

[6] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modeling for early hospital readmission of patients with diabetes in India," in International Journal of Diabetes in Developing Countries, pp. 519-528, June 2016.

[7] Hao, Tianyong, A. Rusanov, and C. Weng, "Extracting and normalizing temporal expressions in clinical data requests from researchers." Smart Health. Springer Berlin Heidelberg, pp. 41-51, 2013.