INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

COMPUTATIONAL PHENOTYPING FROM ELECTRONIC HEALTH RECORDS USING EXTRACTIVE SUMMARIZATION

Eathiraj L[1], Vannamalar I P[2], Vijayarani J[3]

[1]M.E. Scholar, Department of CSE, Anna University, Chennai
[2]Registered Medical Practitioner, M.B.B.S (137977), TamilNadu
[3] Teaching Fellow, Department of CSE, Anna University, Chennai
[1]eathiraj.cse@gmail.com, [2]vannamalar2006@gmail.com, [3]viji.cs66@gmail.com

Abstract: - Text summarization is a process of reducing the complexity of the textual data to focus more on the relevant content by eliminating the irrelevant data. In recent times, the amount of textual data rises on a large scale in clinical domain; which is commonly left unexplored. The textual data gets unexplored in clinical data mining due to the over dependence on images of scans and X-rays for evaluating a patient. In this paper an extractive summarization method is used for summarizing large amount of textual data present in the patient summary records to yield computational phenotypes. Extractive summarization helps in summarizing the contents of a large file by selecting the relevant statements from the data; rather than creating new sentences. It helps to summarize the contents without missing the medical context present in the records which is really significant. In this paper, restricted Boltzmann machine which uses a forward and backward propagation networks is employed for generating the phenotypes. When evaluated with the n2c2 dataset, the proposed model summarizes the electronic health records of patients and generates the phenotypes in an effective manner which will be useful in the field of clinical decision support system.

Keywords: phenotype, summarization, RBM, EHR

1. Introduction

Text data in clinical domain are one of the unexplored areas in biomedical informatics where, natural language processing techniques are used for mining relevant content from the text by normalizing the abundant text present in the electronic health records (EHR). Due to the development of statistical models and recent advances in computational power, it is feasible to do many applications in biomedical informatics. Most of the textual contents present in the EHRs are in the form of unstructured and free text. It is important for the model to summarize the content without losing the contextual meaning of the clinical terms and medical condition of the patients specified in the data.

In this paper, a neural network model is proposed for summarizing the text with less complexity and more accuracy compared to the older models based on rule-based algorithms. Instead of stacking restricted Boltzmann machines (RBM) one over the other, here simple forward-backward propagation technique is used in RBMs for extractive summarization (Verma and Nidhi 2017) of the textual content found in the electronic health records.

Restricted Boltzmann machines learn to reconstruct data by themselves in an unsupervised fashion, making several forward and backward passes between the visible layer and the first hidden layer without involving a deeper network. In the reconstruction phase, the activations of the hidden layer will become the input in a backward pass. They are multiplied by the same weights, one per internal node-edge, and the weight is adjusted in the forward pass. The sum of those products is added to a visible layer bias at each visible node and the output of those operations is a reconstruction of the approximation of the original input.

The objective of this paper is to build an extractive summarization model to summarize the unstructured text in the EHRs. The aim is,

- To produce a summarized record of each patient (such as different tests, different visits and different symptoms), which will be helpful for doctors and clinicians to make good decisions.
- To reduce the complexity found in the unstructured text (i.e free clinical notes) for faster understanding of the patient's condition and for faster decision making.
- To implement a neural network model (such as RBM) for summarizing the electronic health records for better performance compared to the classical rule based algorithms.

Section 1 gives the introduction about text summarization in clinical decision support. Section 2 gives the related work on EHR-based computational phenotyping and visualizing the clinical data. Section 3 describes the overall architecture of the proposed model. Section 4 discusses implementation details and results. Section 5 gives the details about the evaluation. Section 6 concludes the proposed work.

2. Related Work

Sultanum et al. (2019) explained the use of LSTM-CNN for summarization. They proposed an auto encoder and a decoder model for generating summary based on the abstractive summarization. They summarized the text using the curation based approach, which is a more effective way for organizing and selecting the text. However, they had not applied the method for summarizing the clinical text data. Kwon et al. (2018) described the use of visual analytics on textual data for clinical decision support. After performing pre-processing of data, the important keywords are separated from the text and most commonly occurring words are then visualized using a word cloud. However, they had not handled the importance of contextual relevance of text in biomedical informatics in electronic health records. Zeng et al. (2018) discussed various algorithms and methods used in medical phenotype generation. They performed a survey on various deep learning algorithms (Habibi et al. 2017) for generating phenotypes.

3. System Model

This paper aims to build a text summarization model for summarizing the EHR records by extracting features based on nouns and enhancing the features based on scoring of the sentences present in the record by generating sentence-feature matrix. Then this matrix is used for computational phenotyping using an RBM model to produce summaries of EHRs. Then the cosine similarity and Jaccard similarity are used for sorting and arranging sentences for generating an effective summary. In an RBM the hidden units are conditionally independent given the number of visible states. The RBM model used is completely different from the model used for summarizing web documents by Ambekar et al. (2018). It is revised for summarizing EHRs.

Figure 1 describes the overall system architecture of the proposed model for generating computational phenotypes from electronic health records for clinical decision support using extractive summarization. The patient details in the form of EHRs are pre-processed from which features are extracted, vectorized and sentence feature matrices are generated. Algorithm 1 explains the procedure followed in the proposed model.

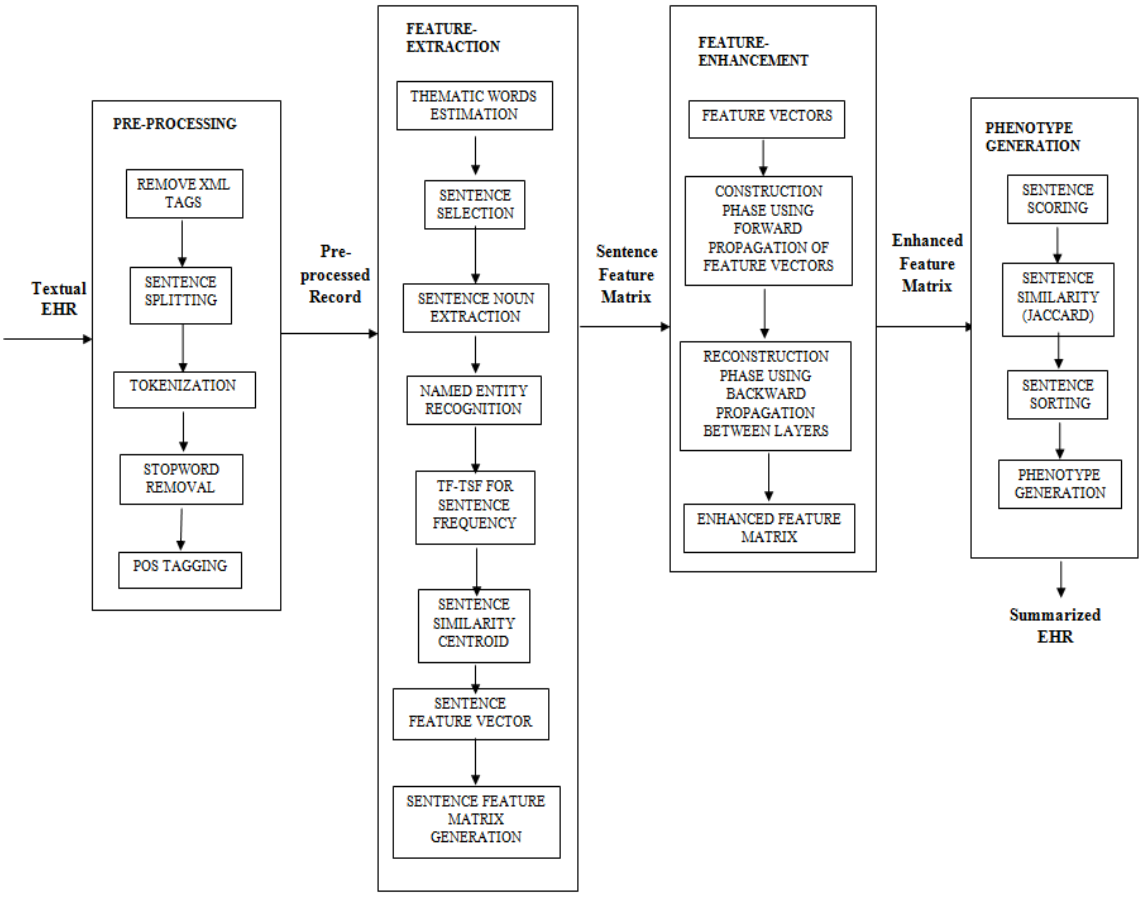


Fig. 1 Phenotype generation using extractive summarization

3.1 Term Frequency-Inverse Sentence Frequency (TF-ISF)

The tf–isf value increases proportionally to the number of times a word appears in the sentence and is offset by the number of sentences in the document that contain the word, which helps to adjust for the fact that some words appear more frequently in general. A centroid sentence is generated for easier sentence organization during the summarization process. The feature vector values are calculated for the given sentences and then a sentence feature matrix is produced by concatenating all sentence feature vectors. For a term **t** in a document d, the weight Wt,d of term t in document d is given by:

$$Wt,d = TFt,d \log (N/DFt) \qquad (1)$$

where TFt,d - the number of occurrences of t in document d
DFt - the number of records containing the term t.
N - the total number of records in the dataset.

```
Algorithm 1
Input: EHR data
Outpt: Phenotype

//Preprocess
for each record (R)
        r1= remove_XmlTags(R)
        r2=sentence_split(r1)
        r3=tokenize(r2)
        r4=stopwords_remove (r3)
        r5=POS_Tagger(r4)
        return(r5)
end

//Feature extraction
for each pre-processed data (r5)
        rf= extract_feature(r5)
        rv=vectorize(rf)
        rfm=generate_sentence_featurematrix(rv)
        return(rfm)
end
//Feature enhancement

for each sentence_feature_matrix (rfm)
        rfrm=sentence_feature_relatiomatrix(rfm)
        return(rfrm)
end

//Phenotype generation

for each sentence_feature_relationmatrix(rfrm)
        assign weight (rfrm)
        sort (rfrm)
        ph=generate phenotype(rfrm) //RBM
        return(ph)
end
```

3.2 RBM algorithm

Restricted Boltzmann machines (RBMs) are effective for individual training of machines for any input given. There is a restriction of communication between the nodes within a single layer. RBMs avoid the need for implementing complex auto-encoders and decoders. Here we use a model to train the weight matrix by using a contrast divergence method. The algorithm performs a Gibbs sampling inside a gradient descent procedure to compute the weight update for sentence matrices. These sentence matrices are sorted using Jaccard similarity. Finally, phenotypes are generated.

3.2.1 Activation function

Each vector is multiplied and added a bias weight in order to achieve an activation function S of input x.

$$S(x) = \frac{e^x}{1+e^x} \qquad (2)$$

Hence, the resultant activation function is,

$$\square^{(1)} = S(v^{(0)T}W + a) \qquad (3)$$

h and v are vectors of hidden and visible layers respectively.

$$v^{(1)} = S(\square^{(1)}W^T + b) \qquad (4)$$

3.2.2 Contrastive Divergence

Boltzmann machines are energy based models and a joint configuration (v,h) of the visible and hidden units has an energy given by

$$E(v,h) = -\sum_{i\in visible} a_i v_i - \sum_{j\in hidden} b_j \square_j - \sum_{i,j} v_i \square_j w_{i,j} \qquad (5)$$

Here vi and hj are the binary states of visible unit i and the hidden unit j, ai, bj are their biases and Wi,j is the weight between them.

Contrastive divergence (CD) is the difference between two Kullback-Leiber divergences.

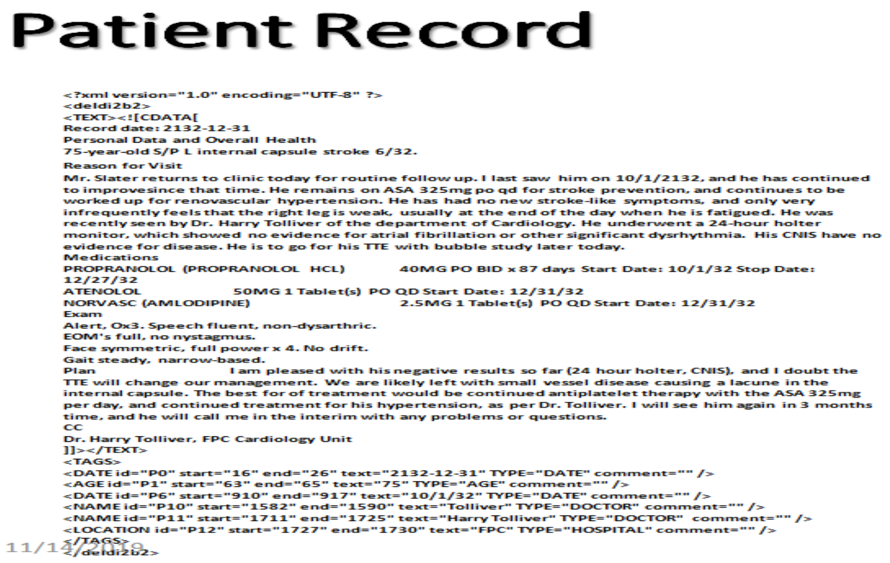
$$CD_k(W, v^{(0)}) = -\sum_{\square} p(h|v_k)\frac{\partial E(v_{k,}, h)}{\partial W} + \sum_{\square} p(h|v_k)\frac{\partial E(v_{k,h})}{\partial w} \qquad (6)$$

4. Experimental setup and results

Dataset n2c2[1] from Harvard university is used for EHR summarization. Xml tags are removed from the data (270 records) using 'termcolor', 'Xml parser' and theano packages. Then the data is split into sentences using 'sentence splitter'. The sentences are then fed into a tokenization algorithm for getting a sequence of individual tokens. Then the most commonly occurring words or stop words are removed to filter out the most occurring words such as “not”, ”the”, ”and”. The thematic words from the given set are extracted. Sentence with highest nouns are given more preference compared to others. Named entities and medical terms are the most important features in clinical text, therefore recognizing them is an important task. For finding the frequency of nouns occurring in the sentences, TF-ISF is used

4.1 Results

An individual patient record is given as an input to the system. The patient record (Figure 2) shows the brief summary of the patient along with the medications and their respective dosages are given.



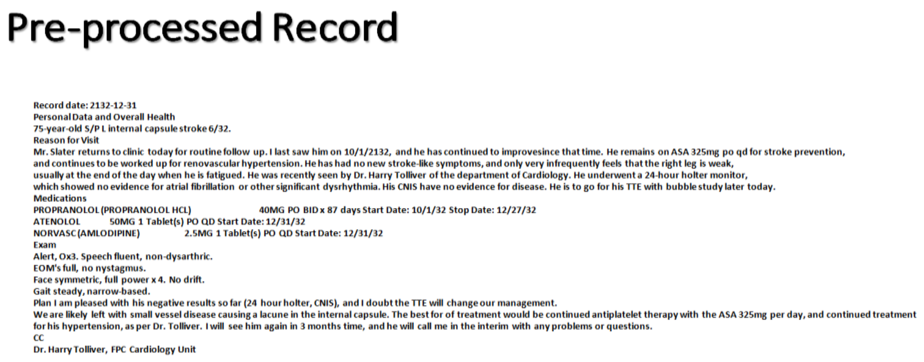
Patient Record

```
<?xml version="1.0" encoding="UTF-8" ?>
<deldi2b2>
<TEXT><![CDATA[
Record date: 2132-12-31
Personal Data and Overall Health
75-year-old S/P L internal capsule stroke 6/32.
Reason for Visit
Mr. Slater returns to clinic today for routine follow up. I last saw him on 10/1/2132, and he has continued
to improvesince that time. He remains on ASA 325mg po qd for stroke prevention, and continues to be
worked up for renovascular hypertension. He has had no new stroke-like symptoms, and only very
infrequently feels that the right leg is weak, usually at the end of the day when he is fatigued. He was
recently seen by Dr. Harry Tolliver of the department of Cardiology. He underwent a 24-hour holter
monitor, which showed no evidence for atrial fibrillation or other significant dysrhythmia. His CNIS have no
evidence for disease. He is to go for his TTE with bubble study later today.
Medications
PROPRANOLOL (PROPRANOLOL HCL)          40MG PO BID x 87 days Start Date: 10/1/32 Stop Date:
12/27/32
ATENOLOL              50MG 1 Tablet(s) PO QD Start Date: 12/31/32
NORVASC (AMLODIPINE)                   2.5MG 1 Tablet(s) PO QD Start Date: 12/31/32
Exam
Alert, Ox3. Speech fluent, non-dysarthric.
EOM's full, no nystagmus.
Face symmetric, full power x 4. No drift.
Gait steady, narrow-based.
Plan                  I am pleased with his negative results so far (24 hour holter, CNIS), and I doubt the
TTE will change our management. We are likely left with small vessel disease causing a lacune in the
internal capsule. The best for of treatment would be continued antiplatelet therapy with the ASA 325mg
per day, and continued treatment for his hypertension, as per Dr. Tolliver. I will see him again in 3 months
time, and he will call me in the interim with any problems or questions.
CC
Dr. Harry Tolliver, FPC Cardiology Unit
]]></TEXT>
<TAGS>
<DATE id="P0" start="16" end="26" text="2132-12-31" TYPE="DATE" comment="" />
<AGE id="P1" start="63" end="65" text="75" TYPE="AGE" comment="" />
<DATE id="P6" start="910" end="917" text="10/1/32" TYPE="DATE" comment="" />
<NAME id="P10" start="1582" end="1590" text="Tolliver" TYPE="DOCTOR" comment="" />
<NAME id="P11" start="1711" end="1725" text="Harry Tolliver" TYPE="DOCTOR" comment="" />
<LOCATION id="P12" start="1727" end="1730" text="FPC" TYPE="HOSPITAL" comment="" />
</TAGS>
</deldi2b2>
```

11/14/2019

Fig. 2 Input patient record

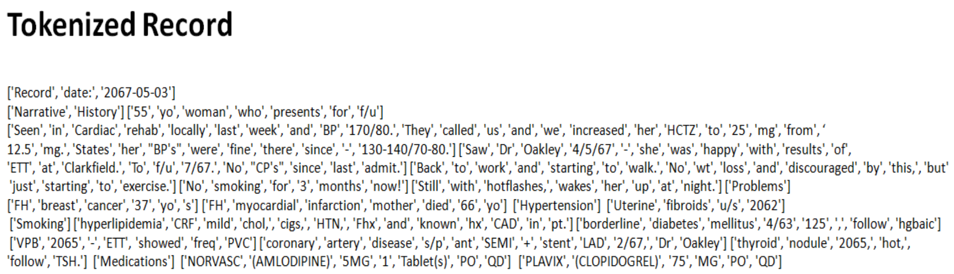
Figure 3 shows a stage of preprocessed output, which results after going through the Xml Tags removal process. In this phase, the medications and their respective dosages are removed as our main area of focus here is textual content present in the records. The output from the tokenization phase (Figure 4) which is a sequence of tokens used as an input for the training stage (Figure 5). The documents are trained individually using the RBM algorithm to produce a sentence feature vector (Figure 6) for the sentences present in the record. The enhanced feature matrix (Figure 7) is generated after feeding the sentence feature vector in the RBM module which is sorted and scored. Figure 8 shows the final phenotype output whose complexity is reduced compared to the original data.



Pre-processed Record

```
Record date: 2132-12-31
Personal Data and Overall Health
75-year-old S/P L internal capsule stroke 6/32.
Reason for Visit
Mr. Slater returns to clinic today for routine follow up. I last saw him on 10/1/2132, and he has continued to improvesince that time. He remains on ASA 325mg po qd for stroke prevention,
and continues to be worked up for renovascular hypertension. He has had no new stroke-like symptoms, and only very infrequently feels that the right leg is weak,
usually at the end of the day when he is fatigued. He was recently seen by Dr. Harry Tolliver of the department of Cardiology. He underwent a 24-hour holter monitor,
which showed no evidence for atrial fibrillation or other significant dysrhythmia. His CNIS have no evidence for disease. He is to go for his TTE with bubble study later today.
Medications
PROPRANOLOL (PROPRANOLOL HCL)          40MG PO BID x 87 days Start Date: 10/1/32 Stop Date: 12/27/32
ATENOLOL       50MG 1 Tablet(s) PO QD Start Date: 12/31/32
NORVASC (AMLODIPINE)          2.5MG 1 Tablet(s) PO QD Start Date: 12/31/32
Exam
Alert, Ox3. Speech fluent, non-dysarthric.
EOM's full, no nystagmus.
Face symmetric, full power x 4. No drift.
Gait steady, narrow-based.
Plan I am pleased with his negative results so far (24 hour holter, CNIS), and I doubt the TTE will change our management.
We are likely left with small vessel disease causing a lacune in the internal capsule. The best for of treatment would be continued antiplatelet therapy with the ASA 325mg per day, and continued treatment
for his hypertension, as per Dr. Tolliver. I will see him again in 3 months time, and he will call me in the interim with any problems or questions.
CC
Dr. Harry Tolliver, FPC Cardiology Unit
```

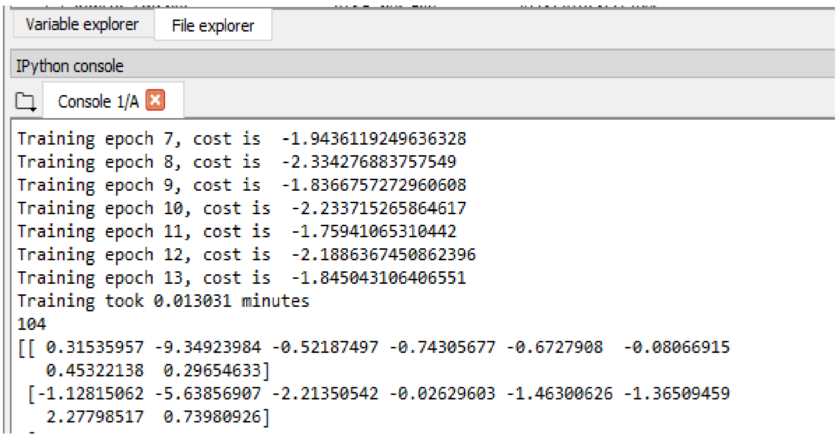
Fig. 3 Pre-processed record



Tokenized Record

```
['Record', 'date:', '2067-05-03']
['Narrative', 'History'] ['55', 'yo', 'woman', 'who', 'presents', 'for', 'f/u']
['Seen', 'in', 'Cardiac', 'rehab', 'locally', 'last', 'week', 'and', 'BP', '170/80.', 'They', 'called', 'us', 'and', 'we', 'increased', 'her', 'HCTZ', 'to', '25', 'mg', 'from', '
12.5', 'mg.', 'States', 'her', "BP's", 'were', 'fine', 'there', 'since', '-', '130-140/70-80.'] ['Saw', 'Dr', 'Oakley', '4/5/67', '-', 'she', 'was', 'happy', 'with', 'results', 'of',
'ETT', 'at', 'Clarkfield.', 'To', 'f/u', '7/67.', 'No', "CP's", 'since', 'last', 'admit.'] ['Back', 'to', 'work', 'and', 'starting', 'to', 'walk.', 'No', 'wt', 'loss', 'and', 'discouraged', 'by', 'this,', 'but'
 'just', 'starting', 'to', 'exercise.'] ['No', 'smoking', 'for', '3', 'months', 'now!'] ['Still', 'with', 'hotflashes,', 'wakes', 'her', 'up', 'at', 'night.'] ['Problems']
['FH', 'breast', 'cancer', '37', 'yo', 's'] ['FH', 'myocardial', 'infarction', 'mother', 'died', '66', 'yo']  ['Hypertension']  ['Uterine', 'fibroids', 'u/s', '2062']
 ['Smoking'] ['hyperlipidemia', 'CRF', 'mild', 'chol,', 'cigs,', 'HTN,', 'Fhx', 'and', 'known', 'hx', 'CAD', 'in', 'pt.'] ['borderline', 'diabetes', 'mellitus', '4/63', '125', ',', 'follow', 'hgbaic']
 ['VPB', '2065', '-', 'ETT', 'showed', 'freq', 'PVC'] ['coronary', 'artery', 'disease', 's/p', 'ant', 'SEMI', '+', 'stent', 'LAD', '2/67,', 'Dr', 'Oakley'] ['thyroid', 'nodule', '2065,', 'hot,',
'follow', 'TSH.']  ['Medications']  ['NORVASC', '(AMLODIPINE)', '5MG', '1', 'Tablet(s)', 'PO', 'QD']  ['PLAVIX', '(CLOPIDOGREL)', '75', 'MG', 'PO', 'QD']
```

Fig. 4 Tokenized record

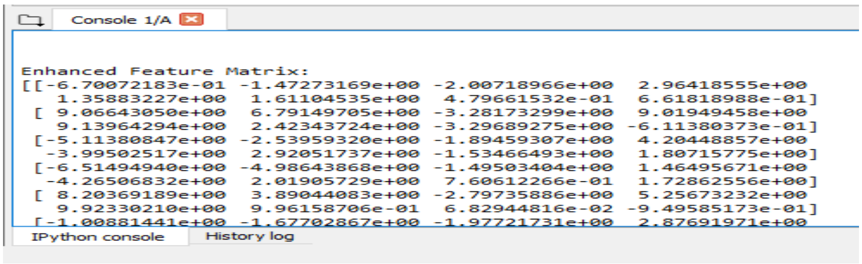


```
Training epoch 7, cost is  -1.9436119249636328
Training epoch 8, cost is  -2.334276883757549
Training epoch 9, cost is  -1.8366757272960608
Training epoch 10, cost is  -2.233715265864617
Training epoch 11, cost is  -1.75941065310442
Training epoch 12, cost is  -2.1886367450862396
Training epoch 13, cost is  -1.845043106406551
Training took 0.013031 minutes
104
[[ 0.31535957 -9.34923984 -0.52187497 -0.74305677 -0.6727908  -0.08066915
   0.45322138  0.29654633]
 [-1.12815062 -5.63856907 -2.21350542 -0.02629603 -1.46300626 -1.36509459
   2.27798517  0.73980926]
```

Fig. 5 Training individual records

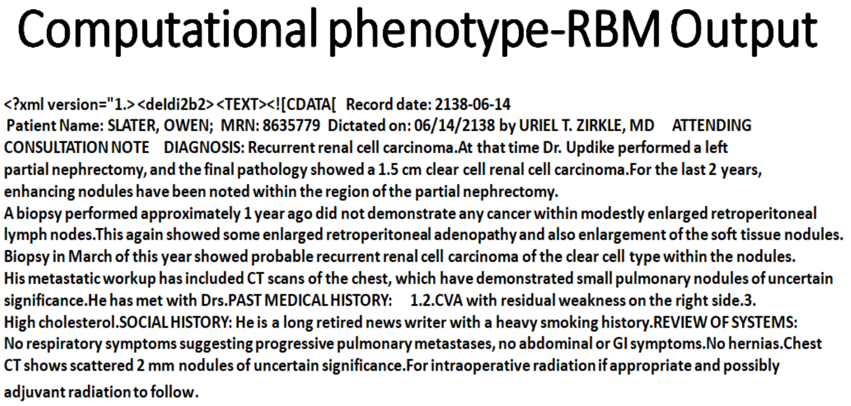
```
[[ 0.31535957 -9.34923984 -0.52187497 -0.74305677 -0.6727908  -0.08066915
   0.45322138  0.29654633]
 [-1.12815062 -5.63856907 -2.21350542 -0.02629603 -1.46300626 -1.36509459
   2.27798517  0.73980926]
 [ 0.55288901  0.82723309  0.83773417 -0.76993612 -2.01195595 -0.1728475
  -0.38878369 -1.20519754]
 [-0.67417853 -2.40852527 -1.07622447 -2.48836758  2.96999743 -0.41047489
   3.08063222 -1.16218236]
```

Fig. 6 Sentence Feature Vector



```
Enhanced Feature Matrix:
[[-6.70072183e-01 -1.47273169e+00 -2.00718966e+00  2.96418555e+00
   1.35883227e+00  1.61104535e+00  4.79661532e-01  6.61818988e-01]
 [ 9.06643050e+00  6.79149705e+00 -3.28173299e+00  9.01949458e+00
   9.13964294e+00  2.42343724e+00 -3.29689275e+00 -6.11380373e-01]
 [-5.11380847e+00 -2.53959320e+00 -1.89459307e+00  4.20448857e+00
  -3.99502517e+00  2.92051737e+00 -1.53466493e+00  1.80715775e+00]
 [-6.51494940e+00 -4.98643868e+00 -1.49503404e+00  1.46495671e+00
  -4.26506832e+00  2.01905729e+00  7.60612266e-01  1.72862556e+00]
 [ 8.20369189e+00  3.89044083e+00 -2.79735886e+00  5.25673232e+00
   9.92330210e+00  9.96158706e-01  6.82944816e-02 -9.49585173e-01]
 [-1.00881441e+00 -1.67702867e+00 -1.97721731e+00  2.87691971e+00
```

Fig. 7 Enhanced Feature Matrix



Computational phenotype-RBM Output

```
<?xml version="1.><deldi2b2><TEXT><![CDATA[   Record date: 2138-06-14
 Patient Name: SLATER, OWEN;  MRN: 8635779  Dictated on: 06/14/2138 by URIEL T. ZIRKLE, MD    ATTENDING
CONSULTATION NOTE    DIAGNOSIS: Recurrent renal cell carcinoma.At that time Dr. Updike performed a left
partial nephrectomy, and the final pathology showed a 1.5 cm clear cell renal cell carcinoma.For the last 2 years,
enhancing nodules have been noted within the region of the partial nephrectomy.
A biopsy performed approximately 1 year ago did not demonstrate any cancer within modestly enlarged retroperitoneal
lymph nodes.This again showed some enlarged retroperitoneal adenopathy and also enlargement of the soft tissue nodules.
Biopsy in March of this year showed probable recurrent renal cell carcinoma of the clear cell type within the nodules.
His metastatic workup has included CT scans of the chest, which have demonstrated small pulmonary nodules of uncertain
significance.He has met with Drs.PAST MEDICAL HISTORY:     1.2.CVA with residual weakness on the right side.3.
High cholesterol.SOCIAL HISTORY: He is a long retired news writer with a heavy smoking history.REVIEW OF SYSTEMS:
No respiratory symptoms suggesting progressive pulmonary metastases, no abdominal or GI symptoms.No hernias.Chest
CT shows scattered 2 mm nodules of uncertain significance.For intraoperative radiation if appropriate and possibly
adjuvant radiation to follow.
```

Fig. 8 Final Phenotype Output

5. Evaluation

ROUGE as well as precision, recall and F-measure are used for evaluating the summarized EHRs. ROUGUE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a software package used for evaluating automatic summarization. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. The following measures are used.

- **ROUGE-N:** Overlap of N-grams between the system and the reference summaries.
- **ROUGE-L:** Longest Common Subsequence (LCS) based statistics. The longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.
- **ROUGE-W:** Weighted LCS-based statistics that favor consecutive LCS.

Precision is known as positive predictive value. It is defined as the number of correct result divided by the number of the retrieved result.

$$\text{Precision} = \sum TP \,/\, (\sum TP + \sum FP) \qquad (7)$$

where,

TP = True positive - total number of positive replies provided by the user which is positive

FP = False positive- total number of negative replies provided by the user which is positive.

Recall is known as a true positive value or sensitivity. It is defined as the number of correct result divided by the number of the relevant result.

$$\text{Recall} = \sum TP / \sum P \qquad (8)$$

where,

TP = True positive - total number of replies which is positive.

P = Total number of replies

F-measure = 2 * Precision * Recall / (Precision + Recall)
(9)

Evaluation with the best and average values (Figure 9) shows the best results are achieved with Rouge-l (R-l) measure. Individual summary of patients are evaluated (Figure 10) prove that better precision, recall and F-measure are obtained for the patient H2R0.

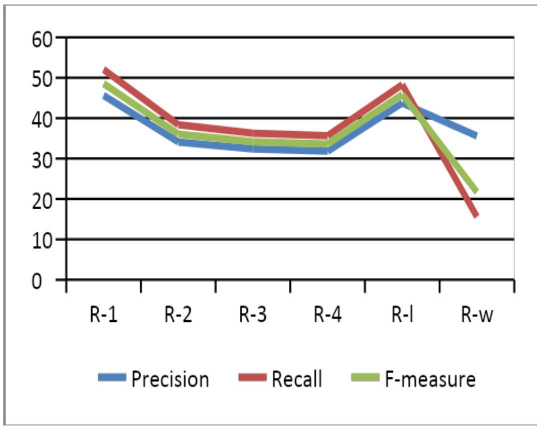


Fig. 9 Evaluation - best and average

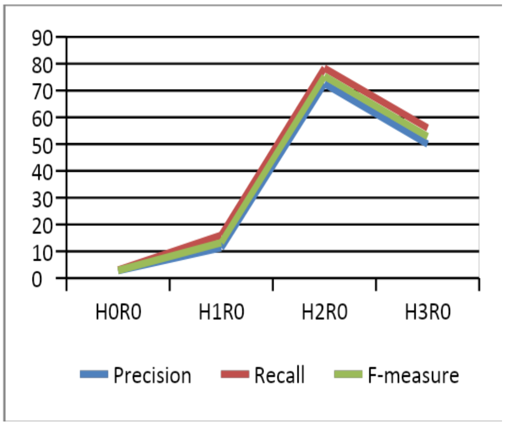


Fig. 10 Evaluation - individual values

6. CONCLUSION

This paper proposed a clinical text summarization model to work on the unstructured text found on numerous patient's EHRs. This model will surely give support to the doctors to avoid mental burnout of themselves and also to aid in clinical decision support for faster and personalized smart healthcare for patients. The individual record summary is produced as a computational phenotype for every patient. It can be clustered and categorized to group the patients for predictive analytics of patients health related conditions. In future, concepts can be extracted from these summaries

REFERENCES

1. Ambekar, A., Shah, K., Agrawal, M., Pawar, S., & Shaikh, A. (2018) Text Summarization Using Restricted Boltzmann Machine: Unsupervised Deep Learning Approach.

2. Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33(14), i37-i48.

3. Kwon, B. C., Choi, M. J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., ... & Choo, J. (2018). Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE transactions on visualization and computer graphics, 25(1), 299-309.

4. Sultanum, N., Singh, D., Brudno, M., & Chevalier, F. (2018). Doccurate: A Curation-Based Approach for Clinical Text Visualization. IEEE transactions on visualization and computer graphics, 25(1), 142-151.

5. Verma, S., & Nidhi, V. (2017). Extractive summarization using deep learning. arXiv preprint arXiv:1708.04439.

6. Zeng, Z., Deng, Y., Li, X., Naumann, T., & Luo, Y. (2018). Natural language processing for EHR-based computational phenotyping. IEEE/ACM transactions on computational biology and bioinformatics, 16(1), 139-153.

A Brief Author Biography

***Eathiraj.L* –** Currently pursuing research in Biomedical Informatics as a postgraduate (M.E) scholar at College of Engineering Guindy. Would like to contribute to the world through a set of skills acquired in computer science into medicine so that the world becomes a smarter and better place for people of all economic and social backgrounds.

***Vannamalar.P* –** Medical Practitioner willing to expand interests in all sorts of technologies which would enhance smarter and personalized medicine.

***Vijayarani.J* –** PhD research scholar at College of engineering guindy Anna University. Have worked in all sorts of projects related to machine translation and natural language processing.