DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION PHI FROM FREE TEXT IN MEDICAL RECORDS

Geetha Mahadevaiah[1], Dinesh M.S[1], Rithesh Sreenivasan[1], Sana Moin[1] and Andre Dekker[2]

[1]Philips Research India, Philips Innovation Campus, Manyata Tech Park, Nagavara Bangalore - 560045, India.
[2]Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Dr Tanslaan 12, 6229ET, Maastricht, The Netherlands

ABSTRACT

Medical health records often contain clinical investigations results and critical information regarding patient health conditions. In these medical records, along with patient health information, patient Protected Health Information (PHI) such as names, locations and date information can co-exist. As per Health Insurance Portability and Accountability Act (HIPAA), before sharing the medical records with researchers and others, all types of PHI information needs to be de-identified. Manual de-identification through human annotators is laborious and error prone, hence, a reliable automated de-identification system is need of the hour.

In this work, various state of the art techniques for de-identification of patient notes in electronic health records were analyzed for their performance, based on the performance quoted in the literature, NeuroNER was selected to de-identify Indian Radiology reports. NeuroNER is a named-entity recognition text de-identification tool developed by Massachusetts Institute of Technology (MIT). This tool is based on the Artificial Neural Networks written in Python and uses Tensorflow machine-learning framework and it comes with five pre-trained models.

To test the NeuroNER models on Indian context data such as name of the person and place, 3300 medical records were simulated. Medical records were simulated by extracting clinical findings, remarks from MIMIC-III data set. For collection of all the relevant Indian data, various websites were scraped to include Indian names, Indian locations (all towns and cities), and Indian Hospital and unit names. During the testing of NeuroNER system, we observed that some of the Indian data such as name, location, etc. were not de-identified satisfactorily. To improve the performance of NeuroNER on Indian context data, along with the existing NeuroNER pre-trained model, a new pre-trained model was added to handle Indian medical reports. Medical dictionary lookup was used to reduce number of misclassifications. Results from all four pre-trained models and the model trained on Indian simulated data were concatenated and final PHI token list was generated to anonymize the medical records to obtain de-identified records. Using this approach, we improved the applicability of the NeuroNER system to Indian data and improved its efficiency and reliability. 2000 simulated reports were used for transfer learning as training set, 1000 reports were used for test set and 300 reports were used for validation (unseen) set.

KEYWORDS

De-identification, Free text, Protected Health Information, Medical records, Radiology reports, Indian context data

1. INTRODUCTION

Clinical documents contain valuable information (patient disease, medical procedure applied and medication) which have resulted in drawing good attention of researchers to explore and extract relevant information from the clinical text, which can have free text, in the form of doctor or nurse notes. However, to use those texts, they have to be de-identified in a way that they give out no personal information on the patient. During the course of PHI identification, it is highly necessary for a de-identification process to retain the medical contents of the records so that this information can help further research and conserve the value of the record [16].

In the US, guidelines for protecting the confidentiality of health care information have been established in the Health Insurance Portability and Accountability Act (HIPAA), which came into effect in April 2003 [7]. Medical records are said to be de-identified when the risk is "very small" that the information can be used alone or in combination with other reasonably available information to re-identify individuals associated with the records. This risk can be estimated and documented statistically for all the medical records in question, or the safe harbor approach can be taken to show that every record is free of 18 specific categories of protected health information (PHI) defined by HIPAA, as detailed in Table 1[8].

PHI Type	Notes
Names	Both full and partial, but not initials
Locations	All geographic subdivisions smaller than a state, including
Dates	All elements of dates (except years) for dates directly related
Ages > 89	All elements of dates (including year) indicative of an age
Telephone numbers	None
Fax numbers	None
Electronic mail addresses	None
Social security numbers	None
Medical record numbers	None
Health plan beneficiary numbers	None
Account numbers	None
Certificate/license numbers	None
Vehicle identifiers	Includes vehicle serial numbers and license plate numbers
Device identifiers and serial numbers	Not restricted to medical devices
Web universal resource locators (URLs)	None
Internet protocol (IP) address numbers	None
Biometric identifiers	Includes finger and voice prints
Any other unique identifying number, code, or characteristic E.g., full face photographic images of full faces, scars or tattoos (and any comparable images).	None

Table 1 PHI Information

Manual de-identification is impractical given the size of electronic health record databases, the limited number of researchers with access to non-de-identified notes [5], and the frequent mistakes of human annotators, so it is quite unfeasible and expensive in terms of time, efforts and

cost[4]. A reliable automated de-identification system would consequently be of high value. Failure to accurately "de-identify" a patient note would jeopardize the patient's privacy. The performance of a de-identification system is therefore critical. In this work, we explored various techniques to identify such information and then de-identify for the purpose of use for researchers [5].

Various state of art techniques for de-identification were analysed for their performance and NeuroNER was selected for further enhancements of de-identification system to address Indian context data [20].

NeuroNER is a named-entity recognition tool based on Artificial Neural Networks written in Python and uses the Tensorflow machine-learning framework. It uses bi-directional LSTM (Long Short Term Memory), along with CRF layer (Conditional Random Field layer). It has five pre-trained models, which are Conll , I2b2 GloVe Spacy, I2b2 GloVe Stanford, Mimic GloVe Spacy and Mimic GloVe Stanford. Among these, Conll was trained on Reuters data which is based on American, European and Asian stock market indices and other four were trained on medical data. These datasets were prepared from various major sources available using SpaCy and Stanford NER taggers. It also uses GloVe pre-trained token embedding.

Since NeuroNER is based on machine learning, its output and efficiency depends upon the kind of data used for model training. In addition, when the tool was further tested, it was observed that it could not perform satisfactorily on Indian data (details are captured in Table 1), as the pre-trained models were not trained on such data.

Proposed work was designed to improve Neuro NER capabilities in following aspects:

- Improve applicability of the system to Indian data
- Improve the efficiency of de-identification system with NeuroNER

2. ORGANIZATION OF PAPER

Section 3 covers state of the art literature and analysis of the gaps in the existing research which has laid the foundations of the proposed methodology.

The section 4 background, describes the Deep Learning model called NeuroNER which is the basis to identify entities of interest in a text.

The section 5 proposed solution, details the analysis of the NeuroNER model and the solution to enhance the model by transfer learning and others techniques, to improve recognition of Indian PHI in a text.

Section 6 Results and Comparative Analysis with existing techniques, explains the results obtained from the proposed methodologies with performance comparisons to existing techniques. Section 7. Discussion and analysis covers the key findings and result analysis with the existing techniques and prove the hypothesis with the recap of final outcome.

Section 8. Conclusion summarizes and provides insights to the usefulness and application of the proposed solution to relevant areas. This section also provides future direction to build on and improve.

3. STATE OF THE ART

In literature, researchers usually follow three standard methods for PHI de-identification. They are Rule based, Machine Learning based and Hybrid methods.

Rule based de-identification systems [9] are based on extensive hand-coded rules and specialized dictionaries. Rule based systems do not require a large amount of training data but different variations have to be captured. Curating rules requires significant manual work. Rule creators make assumptions on the data, thereby limiting flexibility on unseen data.

Machine learning based de-identification systems try to solve the problem by token classification. In literature different machine learning algorithms, including CRFs [3], Support Vector Machines (SVM) [13] have been used. In general ML-based systems perform better than rule-based systems due to the inherent flexibility. ML based systems perform poorly on PHI types which have less data.

Hybrid systems can combine the benefits of both rules and machine learning. Certain PHI types like date are best captured using regular expressions whereas PHI types like name are best captured using machine learning techniques. In [15], a hybrid system combining a token-level CRF, a character-level CRF, and a rule-based classifier was used for de-identification.

In recent years there is a noticeable trend in using hybrid methods which combine rule based system and deep learning networks for de-identification tasks [22]. Among the deep learning network architectures Bi-directional Long Short-Term Memory Networks have been successfully used in the field of Named Entity Recognition [12]. Transfer learning with NeuroNER has been shown to be beneficial for a target set with small number of labels. To the best of our knowledge, we could not find any publication that exclusively covers Indian PHI de-identification.

4. BACKGROUND

Named Entity Recognition: Named-entity recognition (NER) aims at identifying entities of interest in the text, such as location, organization and temporal expression. Identified entities can be used in various downstream applications such as patient note de-identification and information extraction systems. They can also be used as features for machine learning systems for other natural language processing tasks [6, 17]. The main objective is to identify noun phrases or part of noun phrases automatically from the text.

Key design decisions included in NeuroNER system are:

- Chunking and text representation
- Inference and ambiguity resolution algorithms
- Modeling of Non-Local dependencies
- Implementation of external knowledge resources and gazetteers

Named entities are often not simply singular words, but are chunks of text. Therefore, some chunking or parsing prediction model is required to predict whether a group of tokens belongs to the same entity. [14] The output tags are annotated with BIOES (which stand for Begin, Inside, Outside, End, Single) indicating the position of the token in the entity.

NeuroNER: The main components of the NeuroNER are recurrent neural networks (RNNs), in particular, type of RNN called Long Short Term Memory (LSTM).The system is composed of three layers:

- Character-enhanced token embedding layer
- Label prediction layer
- Label sequence optimization layer

The character-enhanced token embedding layer maps each token into a vector representation. The sequence of vector representations corresponding to a sequence of tokens as input to the label prediction layer, which outputs the sequence of vectors containing the probability of each label for each corresponding token. Lastly, the sequence optimization layer outputs the most likely sequence of predicted labels based on the sequence of probability vectors from the previous layer. All layers were learned jointly. [3,5,6,9,11,14,18,19]

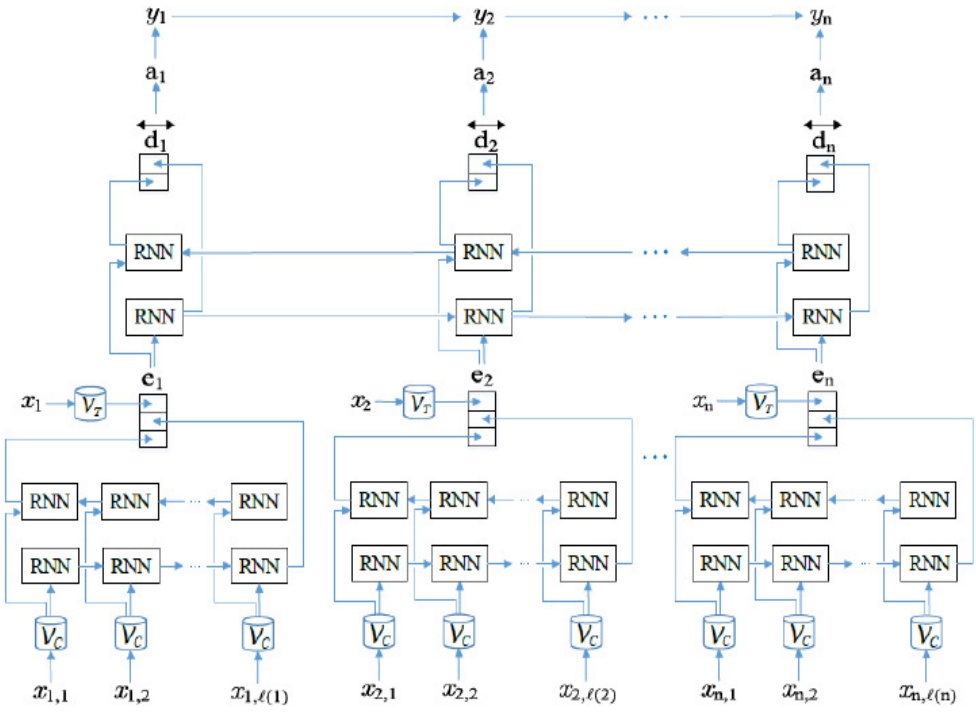


Figure 1 Architecture of NeuroNER model

Figure 1 shows the architecture of the NeuroNER neural network. The type of RNN used in this model is Long Short Term Memory (LSTM). n is the number of tokens, and xi is the ith token. VT is the mapping from tokens to token embedding. l(i) is the number of characters and xi,j is the jth character in the ith token. VC is the mapping from characters to character embedding. ei is the character-enhanced token embedding of the ith token. $\overrightarrow{d_i}$ is the output of the LSTM of label prediction layer, ai is the probability vector over labels, yi is the predicted label of the ith token. [5] .

5. PROPOSED SOLUTION

During the testing with five NeuroNER models, some of the Indian data such as name, location, were not identified by NeuroNER system satisfactorily. After the careful analysis, we proposed to add additional model trained on Indian data to improve the performance of the NeuroNER to address Indian data

5.1. ANALYSIS OF NEURONER

Neuroner's five pre-trained models were used to analyse and test simulated reports with Indian PHI data. During the testing, four models, that is, i2b2 spacy, i2b2 Stanford, MIMIC spacy and MIMIC Stanford shown better results compared to CoNLL since it identified many medical terms as PHI. To maintain efficiency, we excluded ConLL model and used rest of the four models mentioned above for further experimentation. Figure 2 covers the overall flow of Pre-trained models validation of NeuroNER.

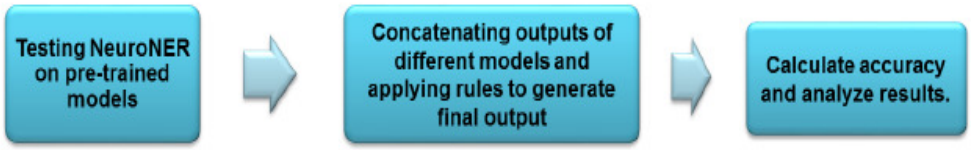


Figure 2. Testing NeuroNER models

5.2. SIMULATION OF REPORTS WITH INDIAN CONTEXT

MIMIC-III data set: The MIMIC-III dataset contains data for 61,532 ICU stays over 58,976 hospital admissions for 46,520 patients, including 2 million patient notes. We used MIMIC de-identified medical reports from MIMIC database. There are various tables present in mimic database, among which we extracted 5000 medical reports from NOTEEVENTS table. PostgreSQL command was used to load and extract data.

Data Extraction: To make a collection of all the relevant Indian data, various websites were scraped, with adherence to their privacy rules, to extract Indian names, Indian locations (all towns and cities), Indian Hospital and unit names. Web scraping is a computer software technique of extracting information from websites. For web scraping, we used BeautifulSoup. It is a python library for pulling data out of HTML and XML files. A parse tree was created for parsed pages that can be used to extract data from HTML.

Data Transformation: After scraping websites, the generated data was thoroughly analyzed and their shortcomings were addressed. Various transformations was performed on data, which includes data cleaning, data formatting and transforming data into a suitable form for experimentation.

Data set creation: Figure 3 shows the process followed to simulate medical reports with the Indian context. Data extracted through web scrapping stored as list were used to replace the anonymized name and other de-identified PHI information that exists in the reports extracted from mimic database. It was also made sure that the simulated reports with Indian context adhere to the format that is suitable for transfer learning with the existing NeuroNER models. Total 3300 reports were simulated. Out of all the simulated reports, 2000 reports were used for training, 1000 reports were used for test and 300 reports were used for unseen validation.

Figure 3. Process to simulate data set with Indian context

5.3. TRANSFER LEARNING

For transfer learning with Indian context data, we used NeuroNER's MIMIC spacy pre trained models and performed transfer learning [1, 21]. During the training process various hyper-parameters such as character based token embedding, LSTM dimension, dropout probability, maximum number of epochs are considered. After the transfer learning, we selected the model with best epoch (epoch 6 was selected out of 30 epochs) and prepared that model to generate labels in de-identification process. During this process we have also introduced new labels which are not standard labels from NeuroNER. To assess the performance, we computed precision, recall, and F-score. Figure 4 covers overall flow of the proposed de-identification process.

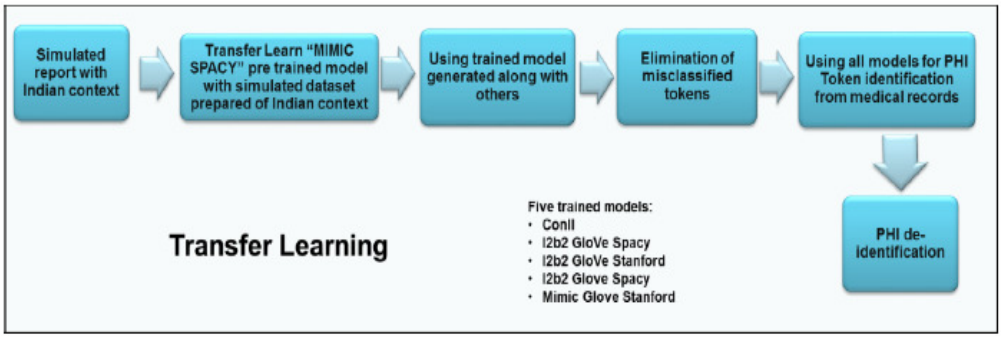


Figure 4. Proposed de-identification model update

We calculate precision, recall and F-score as following:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F-Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (3)$$

TP: True positives, FP: False positives, FN: False negatives, FP: False positives. Intuitively, precision is the proportion of the predicted PHI labels that are gold labels, recall is the proportion of the gold PHI labels that are predicted correctly and F-score is the harmonic mean of precision and recall. GloVe vectors were used for token embedding.

Eliminate misclassification tokens: For reducing the number of misclassifications, we used medical dictionary. We scraped medical website to obtain medical terms and created a hashtable

of it to reduce the access time for dictionary lookup. We check for every token, except person and location, if it belongs to any of the medical term. We label it as not PHI and do not de-identify it.

Concatenation of different model outputs: To improve the efficiency of the system, we developed a technique to concatenate the results from all the pre-trained models. We kept the label-tagging scheme consistent to MIMIC SpaCy type of labels (12 types). We create a text file where we concatenate outputs of all these models according to specifications. For concatenation we used weighted averages where we found certain models had high accuracy for certain PHI information get high weight in the voting when more than one model's result is used for identifying appropriate PHI tag.

PHI de-identification: With all the labelled tags of PHI in report, we anonymize them in the original reports. Every label was anonymized with a dummy value. Dates were shifted by a few years and season were kept intact. Regular expressions are written to mask variables like email, IP address, vehicle number, url and account numbers in case they weren't identified by proposed identifier. After anonymization, we return the de-identified reports as final output.

6. RESULTS AND COMPARATIVE ANALYSIS WITH EXISTING TECHNIQUES

Initially, NeuroNER was tested for its performance on simulated Indian context reports. Table 2 shows the comparison between the claimed result on foreign data and the result obtained when tested on Indian context medical reports.

	Indian Data			Foreign data as reported			
Model	Precision	Recall	F-Score	Model	Precision	Recall	F-Score
i2b2 Stanford	78.97	54.81	64.71	i2b2	97.92	97.84	97.88
i2b2 SpaCy	86.86	60.88	71.59				
MIMIC Stanford	97.39	46.09	62.57	MIMIC	98.82	99.40	99.11
MIMIC SpaCy	97.15	52.01	67.75				
CoNLL	19.85	32.78	24.73	CoNLL	Not Available	Not Available	90.50

Table 2 Comparison of results on different pretrained models

With concatenation of results obtained from i2b2 Stanford, i2b2 SpaCy, MIMIC Stanford, MIMIC SpaCy and MIMIC Spacy India models provides reliable results. On the test set, we processed 1816318 tokens with 45889 PHI and 44398 we correct. Detailed results are captured in Table3.

PHI LABELS	PRECISION	RECALL	F-SCORE	FREQUENCY OF OCCURRENCE IN TEST SET
AGE	93.24%	60.53%	73.4	74
DATE	99.60%	99.56%	99.58	26536
HOSPITAL	95.40%	93.76%	94.57	4757
IDNUM	99.37%	98.51%	98.94	797
LOCATION OTHER	86.75%	93.68%	90.08	1233
NAME	97.06%	96.51%	96.78	10771
PHONE	99.88%	99.94%	99.91	1721

Table 3 Results of each PHI label present in the test set

Table 4 shows the results on validation set after transfer learning on MIMIC Spacy model with simulated Indian data.

	Precision	Recall	F-score
Indian model	97.09	97.19	97.14

Table 4 Retrained model results

7. DISCUSSION AND ANALYSIS

In this work, 3300 reports were simulated by replacing Indian PHI information in the medical reports extracted from NOTEEVENTS table of MIMIC database. Out of all the simulated reports, 2000 used for training, 1000 used for testing and 300 used for unseen validation. During the testing of NeuroNER's models on simulated data, four models, that is, i2b2 spacy, i2b2 Stanford, MIMIC spacy and MIMIC Stanford shown better results compared to CoNLL since it identified many medical terms as PHI. To maintain efficiency, we excluded ConLL model and used rest of the four models mentioned above for further experimentation. For transfer learning with Indian context data, we used NeuroNER's MIMIC spacy pre trained models and performed transfer learning. With the transfer learning, MIMIC Spacy model on simulated Indian PHI data shows an improvement in F-score from 67.75 to 97.14. Results obtained from concatenation of NeruoNER's four models and model trained on Indian simulated data provides significant improvement in the processed tokens as shown in table 3.

8. CONCLUSION

To improve the performance of text de-identification on Indian PHI data, NeuroNER's Deep Learning pre-trained models were updated with transfer learning on simulated data. Since Indian community is spread across the world, proposed approach can be extended to different English speaking countries to de-identify medical reports. In Deep Learning, more data is merrier, updating models with more data is always in scope to improve PHI de-identification performance. For further improvements, different Deep Learning architectures can be explored.

REFERENCES

[1] Andrew Arnold, Ramesh Nallapati and William W. Cohen. Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition". Proceedings of ACL-08: HLT, 2008

[2] Bin He, Yi Guan, Jianyi Cheng, Keting Cen, and Wenlan Hua. 2015. "Crfs based de-identification of medical records." Journal of biomedical informatics 58:S39–S46.

[3] Bui, D. D. A., M. Wyatt, and J. J. Cimino, "The UAB informatics institute and 2016 CEGS N-GRID de-identification shared task challenge", Journal of Biomedical Informatics, 2017.

[4] Dorr DA, et al: "Assessing the difficulty and time cost of de-identification in clinical narratives". Methods Inf Med 2006, 246-52.

[5] Franck Dernoncourt, Ji Young Lee, Peter Szolovits, Ozlem Uzuner. "De-identification of Patient Notes with Recurrent Neural Networks." arXiv:1606.03475v1 [cs.CL] 10 Jun 2016

[6] Franck Dernoncourt, Ji Young Lee, Peter Szolovits. "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks." arXiv:1705.05487 [cs.CL] 16 May 2017

[7] GPO, U.S: 45 C.F.R. § 46 Protection of Human Subjects 2008 [http://www. access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html]

[8] GPO, U.S: 45 C.F.R. § 164 Security and Privacy 2008 [http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html].

[9] Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. "Automated de-identification of free-text medical records." BMC Medical Informatics and Decision Making, 8 (1) (2008), pp. 641-717, 2008 8:32, PMC2526997 2008

[10] Ji Young Lee, Franck Dernoncourt, Peter Szolovits, "Transfer Learning for Named-Entity Recognition with Neural Networks" arXiv:1705.06273 [cs.CL]

[11] Ji Young Lee, Franck Dernoncourt, O¨zlem Uzuner, Peter Szolovits. "Feature Augmented Neural Networks for Patient Note De-identification." arXiv: 1610.09704 [cs.CL] 30 Oct 2016

[12] Kaung Khin, Philipp Burckhardt, Rema Padman, "A Deep Learning Architecture for De-identification of Patient" Notes: Implementation and Evaluation (arXiv:1810.01570 [cs.CL])

[13] Khalifa, A. and S. Meystre, "Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes", Journal of Biomedical Informatics 58 (Supplement), S128-S132, 2015.

[14] Lev Ratinov and Dan Roth. "Design Challenges and Misconceptions in Named Entity Recognition." CoNLL '09 Proceedings of the 13th Conference on Computational Natural Language Learning, Stroudsburg, PA, 2009, pp. 147-155.

[15] Liu, Z., B. Tang, X. Wang, and Q. Chen "De-identification of clinical notes via recurrent neural network and conditional random field" Journal of Biomedical Informatics 75, S34- S42, 2017.

[16] Meystre et al.: "Automatic de-identification of textual documents in the electronic health record: a review of recent research". BMC Medical Research Methodology 2010 10:70

[17] Morrison FP, et al: "Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes?" J Am Med Inform Assoc 2009, 16(1):37-9

[18] Özlem Uzuner, Yuan Luo, Peter Szolovits; "Evaluating the State-of-the-Art in Automatic De-identification." Journal of the American Medical Informatics Association, Volume 14, Issue 5, 1 September 2007, Pages 550–563

[19] Shweta, Asif Ekbal, Sriparna Saha ,Pushpak Bhattacharyya. "Deep Learning Architecture for Patient Data De-identification in Clinical Records." ClinicalNLP@COLING 2016

[20] Szarvas G, Farkas R, Kocsor A: "A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms". 9th Int Conf Disc Sci (DS2006), LNAI 2006, 267-278

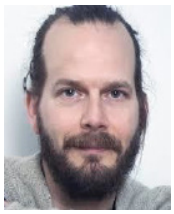
[21] Szarvas G, Farkas R, Busa-Fekete R: "State-of-the-art anonymization of medical records using an iterative machine learning framework". J Am Med Inform Assoc 2007, 574-80

[22] Vithya Yogarajan, Michael Mayo, "A survey of automatic de-identification of longitudinal clinical narratives", Bernhard Pfahringer "arXiv:1810.06765 [cs.AI]"

AUTHORS BIOGRAPHY:

Geetha Mahadevaiah, (Corresponding Author) Senior Director at Philips, Research Department, PIC, Bangalore.30+ years of experience in software engineering and management. Areas of interest : Clinical decision support systems, Semantic Web, healthcare applications Bachelor of Engineering in Computer Science and Technology Bangalore University Master of Business Administration, Bangalore University

Dinesh M.S. Senior Principal Scientist at Philips, Research Department, PIC, Bangalore.19+ years of post-doctoral experience in applied research (healthcare domain). Areas of interest: Machine Learning, Pattern recognition and Image Processing Education: Bachelor of Engineering, Master of Technology and Doctor of Philosophy from University of Mysore



André Dekker is a professor of clinical data science at Maastricht University and has been leading the development of prediction models in radiation therapy for many years. He is also coordinator of t he Personal Health Train project, aiming to facilitate citizen science. Areas of interest : Radiomics, Semantic Web, Radiotherapy, machine learning.