Comprehensive Review of Deep learning Techniques in Electronic Medical Records

[1]S. Biruntha, [2]M Revathy, [3]Raashma Mahaboob, [4] V Meenakshi

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India. Email – banupriya12317@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, Kalaignar Karunanidhi Institute of Technology, Coimbatore, India. Email - revathymkit@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Karpagam University, Coimbatore, India. Email – raashma.mahaboobkahedu.in
[4] Associate Professor, School of Electrical and Electronics Engineering, Sathyabama Institute of Science and Technology, Chennai, India. Email – meenagovind1@yahoo.in

Abstract. A digital collection of patient's health care data likediagnosis history of patient, treatment details, medical prescriptions are stored electronically. This electronic patient health records (EPHR) model provides huge volume of real time data and used for clinical research. Natural Language processing (NLP) automatically retrieve the patient's information based on decision support system. NLP performs traditional techniques of machine learning, deep learning algorithms and focussing on word embeddings, classification and prediction, extraction, knowledge graphs, phenotyping,etc. By using NLP technique, extract the information from clinical data and analysis it provides valuable patient medical information. NLP based on clinical systems are evaluated on document level annotations which contains document of patient report, health status of patient, document section types contain past medical history of patient, summary of discharge statement, etc. similarly the semantic properties contain severity of disease in the aspects of positivity, negativity. These documents are developed and implemented on word level or sentence level. In this survey article, we summarize the recent NLP techniques which are used in EPHR applications. This survey paper focuses on prediction, classification, extraction, embedding, phenotyping, multilingually etc techniques.

Keywords: *Electronic Health Record, Medical Information,NLP, EPHR, document level, deep learning, machine learning.*

I Introduction

In recent years, usage of digital data is rapidly increases and it is necessary to use these data in various research filed work. The digital collection of patients health care data, past medical history of data, treatment details, observation of patient's health status etc are stored electronically called as Electronic Patient Health Record (EPHR) [1]. Data which are stored in EPHR are in various formats like structured data and unstructured data [2].Structured EPHR data contains heterogeneous data likemedication, diagnosis of disease etc. In the unstructured data contains discharge summaries, medical notes etc [3]. In the NLP, analysis of text data processing, extracting the clinical text information from different applications

such as data mining, prediction of various chronic diseases, analysis ofdiseases with its side effects etc. NLP based on deep learning, machine learning techniques produces better performance in the field of health care and biomedicine of clinical text information [4]. In the patient's heath record-based information, NLP is a field of research in the analysis of medical informatics, bio-medicine and computer-basedlinguistics [5].

Electronic Patient Health Records (EPHR) contains the main source of data about patient's treatment details, assessment of diagnosis of disease, patient's health history in the clinical care of the patient. This main source of data in the EPHR includes both structured and unstructured data. Thepatient's information includes laboratory test results, sign for disease, procedures to be conducted in the analysis of disease are considered as structured data. As well as observation of health issues of patient, planning for treatment are considered as unstructured data in the clinical care [6]. In the predictionof disease from the data stored in EPHR which are mainly focused on structured data and detecting the presence or absence of disease from the clinical care [7]. For identifying the health status of patient, the structured data alone is not enough. Along with structed data, patient's age, past history and some other related information is needed for patients'health care [8].

Clinical care NLP requires an automated extractionof patient information system for detecting the disease is an essential one. This automated system only extracts the status of disease, medication from the clinical care. Consequently, pre-processing information is required for extracting the information because the decision-support systems and summarization cannot be performed based on its input data. Therefore, pre-processing includes analysis of structure of the document by tokenization, sentence splitting, spell checking, parts of speech tagging, Word Sense Disambiguation and soon [9]. There are different extraction techniques are available in the clinical care-based NLP model. They are pattern matching techniques, machine learning, deep learning, statistical techniques and rule-based techniques. By using the extracted information analysis, the clinical health care of patient as well as improve the automated decision supported system [10].

The paper has been organized as follows: Section 2 discusses about preliminaries section 3 describes various tasks in health care using NLP, Section 4 discusses methods inhealth care using NLP, Section 5 about applications of NLP in health care domain, Section 6 discuss about framework ofNLP, Section 7 discusses challenges in EPHR, Section 8 explains datasets in FSA, Section 9 concludes the paper with future works.

2. Preliminaries

A) Overview of Natural Language Processing (NLP) Combination of computer science, concept of artificialintelligence (AI) and linguistics are sub field of NLP. Themain aim of NLP is interpreting the human understandablelanguage and analysis it to execute various tasks liketranslation of language, automatic question and answering etc.in handling the different complexities of NLP includes utilizing linguistic, understanding, global, socialistic information, semantic analysis, syntactic analysis,morphological analysis and tokenization. NLP is a branch of AI which focussing on perception of human-generated text data, human-generated speech data. There are various subdisciplines of AI concept in the NLP includes Natural Language Query (NLQ), Natural Language Understanding (NLU), and Natural Language Generation (NLG) [11]. Overview of NLP is given in Figure 1.

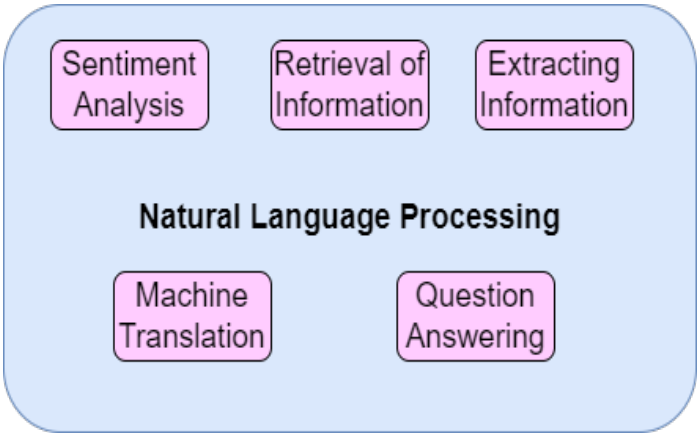

Fig 1 Overview of NLP

Figure 1 shows that overview of NLP contains sentiment analysis, Retrieval of information, extracting information, machine translation and question answering.

1) Sentiment Analysis

In the natural language processing Sentiment analysis, also known as opinion mining, and it extract the subjective textual information. The purpose of sentiment analysis is to determine the strength of sentiment in the textual information and it is used for decision making process. The sentiment is determinedas subjectivity polarity and strength in polarity. The subjectivity polarity is determined as positive or negative. Similarly for strength in polarity is determined as strongly positive, mildly positive, weakly positive etc. for a review the text document.

2) Information Retrieval

Based on the given query it retrieves the relevant information from the dataset. An information retrieval model predicts and searches the collection of natural language document and retrieves the set of matching documents based on the user's query text information.

3) Extracting information in NLP

In the NLP extracting information plays powerful concept and it will enable to parse through textual information. The textualdata contains huge amount of information but for theprocessing all information are not needed. Therefore, it isnecessary to extract the specific textual information and make the relationship between the textual data.

4) Machine Translation

In NLP, machine translation (MT) is one of the components which converts text or speech from one natural language into another and at the same time it preserves the meaning of input text and produces the output of the language. The MT issubfield of artificial intelligence.

5) Question Answering

It focuses on building a model that automatically answer the questions asked by humans in a natural language.

6) Natural Language Processing in Healthcare Domain

National Language processing (NLP) faces many challenges and issues in related with healthcare domain. Huge amount of digital information related with healthcare domain is available in the internet and the digital information includes e-health records, publications based on medicine and symptoms, treatment for the diseases. There are many critical issues in the aspects of digital information are in the form of textual format not in the pre-structured format. It is tedious to enter the patient health history in a text format, controlling, and using of information in the health research field. Therefore, NLP is needed which converts the textual information of health care domain into structured format and it can be used in the computer applications [12]. After that, all medical information is in the structured format, which is easy to ease, time saving and also cost reduction. This medical information is stored electronically heath care of records (EPHR). Physician also practices in the adoption of EPHR because of the act followed from 2009 and the act is Health Information Technology for Economic and Clinical Health (HITECH) Act [13]. The basic format of EPHR is used in 84% of hospitals which has tremendous increased 9-fold from 2008 and based on the recent research from and get the study from office of the National Coordinator for Health Information Technology (ONC). Furthermore, the usage of EPHR by office-based physician has increased from 44% to 88%. The information about patient details is stored in the EPHR model includes diagnosis of disease, demographic information, laboratory reports, drugs, clinical notes, radiological images, etc.

The usage of EPHR has increased in both hospital and outpatient care of data. The main benefits of using EPHR in the healthcare domain enhancing the patient care, minimize the error rate, improving the treatment quality in an efficient way and also it provides rich of medical information to the researchers [14]. In the EPHR model various types of information are stored which includes diagnosis of disease, demographic information, laboratory reports, drugs, clinical notes, radiological images, etc. it is difficult to handle the information like numerical quantities, time series of data, date/time related information and so on. This complicated information in the EPHR is given below:

a. Numerical form of Quantities

This numerical quantity of data includes body mass index value.

b. Date/Time information

This information is related with patient's date of birth and patient's admission date and time details.

c. Natural Language Free Text of information

This information includes heath status of the patient, dischargesummary etc. This information is collected and stored in the EPHR chronologically order.

3 Literature Survey

This section presents survey on various techniques used in the EHR analysis. It also shows how NLP play vital role in EHR practice. Table 1 presents survey on various papers in the EHR data extraction.

Table 1. EHR Survey

Paper	Research area	Technology Used	Limitation studies
Martin et al[1]	cardiovascular clinical research	embedded pragmatic (post marking) studies	It cannot be possible for large collectionof data.
Cao Xiao et al[3]	Review on EHR	Deep learning models	-----
Hasan et al [4]	EHR	Deep learning with NLP	Deep learning applied to solve NLP task.challenges not discussed
Savova et al[5]	Text extraction from EHR	Review on various NLP,data extraction models	The biomedical test records are different from HER. challenges on extractingbiomedical data.
Homaet al [6]	Diseases status	Cognitive techniques	There is less accurate result. Need ofNLP and ML techniques to improve prediction.
Yang et al [7]	Diseases predisction	Text mining approaches	The text mining does not support somecategorical data.
Kirk et al [8]	EHR identification and classification	Machine learning (SVMand conditional random fields)	Assertion and feature selection needsstrong textual analysis using NLP.
Demner et al [9]	NLP in EHR	NLP techniques	Review analysis
Essam et al [10]	ML for clinical data analysis	Review	--------
Knake et al [12]	Pre term health data analysis	ML extractiontechniques	Review analysis
Casey et al [13]	social/behavior of HER analysis	Environmental, socialand health determinants	Imaging data is not studied
Finiegula et al [23]	Biomedical	Entity recognition models	The data analysis has more predictionerror
Shinyama et al [24]	Language analysis	Human languageprocessing	Less efficiency and less robustness
Rink et al [25]	EHR clinical analysis	Automatic data extraction	Semantic analysis is missing.
Si et al [27]	Cancer analysis	ML and NLP textextraction	Accuracy can be improved
Xu et al [29] and Aramaki et al [30]	Drug analysis	Drug label extraction and clinical drug analysis	Need of more efficiency in text analysis
Bethard et al [32]	Clinical data extraction	Semantic data analysis	The run time of analyzing single text takemore time

The EHR analysis requires NLP and deep learning models for effective data analysis and classification. Recently artificialintelligence and deep learning techniques play vital role in

feature extraction and classification for best outcome.

4 Task in Healthcare Using NLP

In the clinical health care domain for the patient based on NLPincludes several tasks and are - Detection of adverse in drug Events (DADEs), extracting the information (EI), Recognitionof name entity (RNE), Clinical Relation Extraction (CRE), Disambiguation of Word Sense concept.

A) Detection of adverse in drug Events (DADE)

Detection of adverse drug events based on intervention of medical treatment through medicines like wrong diagnosis, mistake in prescription, over dosage, medicine allergic reactions etc [15].

Detection of adverse drug events is benefited by the research in the medical field and medical treatment given by the hospital. The information contained in the EPHR are hidden the unstructured data like clinical notes, medical history of the patient, treatment procedure notes, discharge summaries, and testing report form the laboratory [16-18].

Identification and detection of information related to DADE from the clinical notes is a difficult task and consumption of time is high. Therefore, NLP is needed which helps to developautomatic model based on the concept of EPHR for the detection of drug activities, DADE with its interaction to the patient [19].

B) Extracting the Information (EI)

It is very essential in health care domain and NLP uses the EPHR for taking decision of clinical health care support, improving the quality and doing the research basedon the clinical information. In the medical domain EI also automatically extract and encode the clinical information fromthe clinical notes. But in the case of general domain this EI is used to recognized a specialization area in NLP and automatically extract the concepts, events, entities along with its relation between the attributes from textual data [20].

C) Recognition of name entity (RNE)

It is a subtask of extracting information (EI) and also it plays a vital task in the field of health care domain-based NLP. It converts the unstructured textual information into structured textual information and easily readable by the computer [21]. The aim of RNE task is identifying the expressions in the structured textual information whichdenotes the entities like medications, lab tests, and diseases from the clinical notes. Many techniques used in the RNE usesdeep learning, rules model, dictionary model, hybrid approach and statistical approach [22,23].

D) Clinical Relation Extraction (CRE)

It is also a subtask of extracting information (EI)which focuses on identification and detection of semantic relationship between clinical treatment-based concepts from clinical notes [24,25]. For example, from the clinical notes, thetest report of MRI for the patient which reveals that cordcompression in C5-6-disc herniation. This report tells that the patient is affected by two types of diseases like cord compression and C5-6-disc herniation. Here it relates with oneanother by clinically notes. This is the relation of one disease with another by taking single report. Thus, various types of relations are revealed by the existing researchers such as attribute of disease by pairing extraction, identification of temporal

relation, and detection of adverse in the drug event etc [26-30]. NLP in the clinical domain shared various tasks from the clinical notes like Integrating Biology and the Bedside (i2b2) challenges [31], challenges in SemanticEvaluation (SemEval) [32].

E) Disambiguation of Word Sense concept

In the NLP, Word sense disambiguation is the process of knowing meaning for certain word which is activated by particular context. That is, it automatically determines the accurate meaning for the particular context of data. In the health care domain, NLP task requires accurate meaning for ambiguous word. The list of all possiblemeanings for the health care domain is generated based on Word sense disambiguation concept in the NLP. Syntactic or semantic ambiguity, Lexical ambiguity are the problems in NLP. To solve the problem in resolving semantic ambiguity is termed as Word sense disambiguation. Resolving semanticambiguity is difficult than resolving syntactic ambiguity. To solve the problem in word's syntactic ambiguity is by using Part-of-speech (POS) tag in an accurate way [33-37].

5 Methods in Healthcare Using NLP

In the EPHR, extracting the medical information withstructured clinical data and it involves description of structured textual clinical data. There are deep learning (DL), machine learning (ML) and rule-based techniques are exploredin clinical domain. Machine or deep learning model employed features with appropriate algorithms. Figure 2 shows that comparison of EPHR domain applying above concepts with number of publications per year.

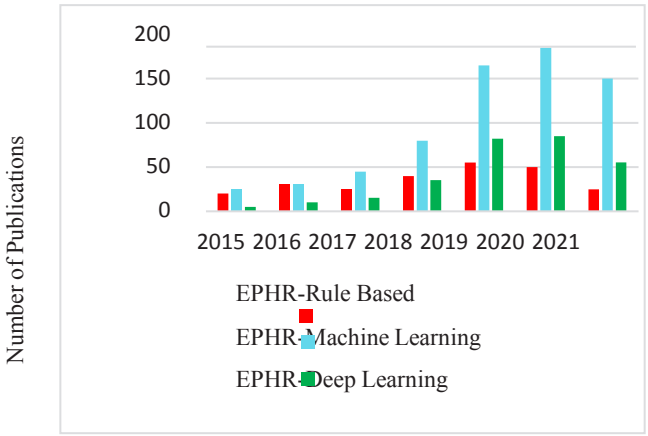

Fig 2. Comparison of various algorithmic concepts

In the observation of figure 2 describes that machine learning algorithm is better growing compared with deep learning/rule-based algorithm. The efficiency of machine learning algorithm is highlighted and is compared with machine learning algorithms [38]. In the recent years, health care domain in the NLP based researchers has shown that deeplearning algorithm plays a vital concept. In the evaluation of biomedical text information Recurrent Neural Network (RNN) with recognition of name entity (RNE) task produces better performance in an effectively. They proposed a model with a combination of Bidirectional Long Short Term Memory (Bi- LSTM) with Condition Random Field (CRF) based on the concept of word-embedding in character level[39]. Habibi et al. [40] proposed a model with a combination of BiLSTM- CRF and it is implemented by Lample et al. [41]. The concept of word embedding is developed by Pyysalo et al. [42]. Comparing these models word

embedding in character level produces high successful rate in the implementation of health care domain-based NLP.

A) Rule-Based Approach

Rules are applied to the textual data and 8ocusing on pattern matching or parsing in the document. These rules are defined in words or POS tag as regular expressions. The steps involved in rule-based approach are shown in Figure 3.

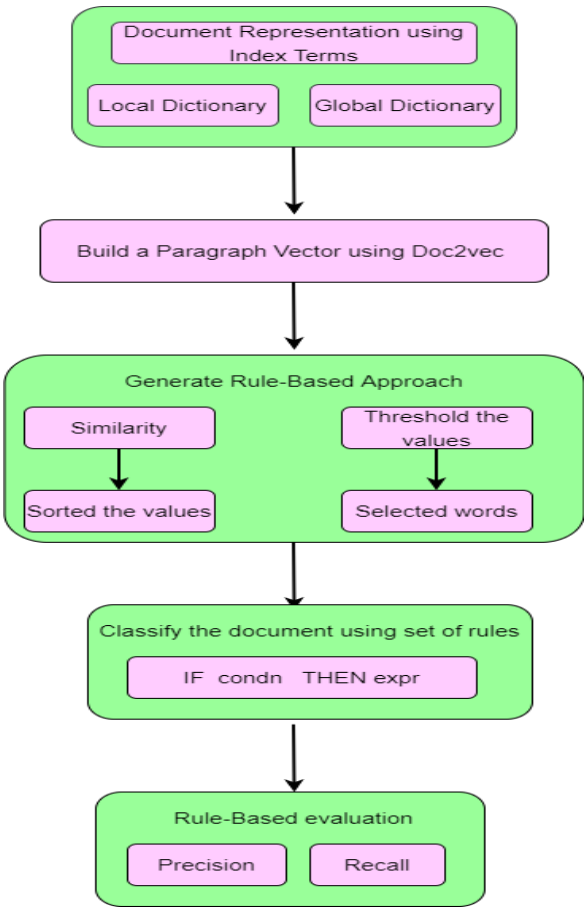

Fig 3. Steps in Rule-Based Approach

From the Figure 3, it describes that step involved in rule-based approach in NLP. In some model riles areconsidered as pattern. Based on the pattern value it represents the text document using index term value. Then by usingDoc2vec it converts the textual document into vector. Document is classified using classifier of IF. THEN is the pattern of the document. The rules are generated in two ways. They are manual methods and automatic method from the dataset. In the evaluation of rule-based approach is implementedby performance metric measures of precision and recall. Rule based approach provide high precision but low recall since the rules for a specific dataset cannot be created for other data sets[43].

B) Machine Learning Approach

In analysis the EPHR in the NLP platform, machine learning algorithm is a sub field of artificial intelligence and it plays a vital role in processing EPHR. This machine learning techniques has two categories like supervised and unsupervised learning. The function of Supervised learning technique is represented by

$$p = f(q) \quad (1)$$

Here the input q provides the mapping function to output p. The algorithms which are used in the supervised model are classification and regression. The commonly used machine learning technique is logistics regression (LR) and support vectors machine (SVM). Similarly for the unsupervised techniques are principal component analyses (PCA), and cluster analysis. In the unsupervised learning technique is learn about the input $q$distribute its features. This input distribute features are considered as set of attributes which are extracted for every datapoint. In recent years for the processing of EPHR uses SVM, LR and random forest (RF) were used [44]. New modern NLP platforms are refined using these innovative machine learning techniques. They are composed of four steps like models, data, loss functions and an algorithm [45-46].

From the figure 4, shows that three layers namely input, hidden and output. In some situations, more than one hidden layer is constructed. It depends upon the situation of problem or data used in constructing the architecture [47]. In the training data set based on the multilayer representation of deep learning techniques are implemented automatically with several factors of unlabeled data. Along with it implements computing resources in GPU, new frame work of algorithm is adapted [48]. Deep learning architecture uses ANN concepts. Recurrent network architecture is unsupervised hierarchicaldata representation.

As from the Figure 4, this ANN architecture contains multiple interconnected nodes (neurons) organized in threelayers: input, hidden and output. In the training process the weight value is updated in the between input to hidden layer 1 and also between hidden layer 1 and layer 2. Finally interconnect with hidden and output layer. In order to get a optimized output of ANN architecture is evaluated byminimizing the loss of function which is defined as:

$$F(\theta, M) = -\sum_{t=0}^{M} \log P\,[Y = q_t|p_t\,, \theta] + \lambda\|\theta\|_y \quad (2)$$

C) Deep Learning Model

It is a subfield of machine learning based on the multi‡ayer neural network architecture. Figure 4 shows that thearchitecture of deep neural network. This multiple layer is used to store the hierarchical representations of data.

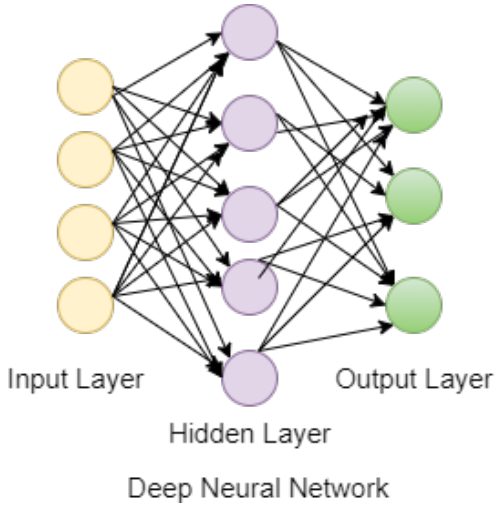

Fig 4. Architecture of Deep neural network

In the training data set M is minimized by applying the log loss function of the first term. In minimizing the second term by using the learned parameter of θ^t with tunable parameter of λ. By implementing this it will prevent thesystem from overfitting and enhancing the model, the back propagation technique is used for minimize the loss function and optimized the final layer loss of the function [49]. The most common deep learning techniques which are used in the NLP concept for health care domain are Multilayer perceptron (MLP), Convolutional neural networks (CNN), Recurrent neural networks (RNN), Restricted Boltzmann machine (RBM).

6 Applications of NLP in Health CareDomain

In the EPHR based on the NLP platforms used both machine learning and deep learning techniques in many ways. Table 2 Shows that application of machine learning approaches in NLP platform with health care domain.

Table 2. Application of Machine Learning approach

Machine Learning Model	**NLP platform in Health care Domain**
Naïve Bayes[50-56]	Heart disease prediction
	Detection of multiple sclerosis
	Cancer and obesity classification
SVM [57-59]	Heart disease prediction
	Diabetes
	Analysis of breast radiology report
Conditional random fields[60-62],	Heart disease prediction
	Diabetes
	Detection of multiple sclerosis
	Analysis of breast radiology report
	Detecting Tumour
Random Forest [63,64]	Heart disease prediction
	Tumour detection
	Classification of cancer

Suliemanet.al [71] effectively classifies the patient portal messages using CNN. And for the biomedical text for recognising the named entities in the CNN is applied in NLP platform [72].

In the observation of Table 2, machine learning algorithm is used in health care domain in various forms. The drawback of using machine learning algorithm is difficult to handle complex high-scale data, even though it is easy to use, simplicity, interpretability in EPHR. To overcome this drawback deep learning algorithm is used. Features of deep learning algorithm is hierarchical in construction of data features and efficient in handling long-range of data dependencies. In the health care domain-based on EPHR research is done in numerous projects using deep learningalgorithm. Also, it provides enhanced result and less consumption time. Deep learning models like CNN, feed forward NN (FFNN) and RNN that can be applied in theanalysis of EPHR [65]. In the pre-processing Vector- embedding technique is applied and also transfer learning has improved its performance in the model [66]. CNN algorithm produces an effective performance in health care domain inthe platform of NLP. S. Baker.et al [67] proposed that identification of cancer using CNN. Y. Peng et.al [68] identified the protein-protein interaction relations in the biomedical report. M. Asada et.al [69] uses CNN and implements the mechanism for extract drug-drug interactions in the NLP platform. M. C. Chen et.al [70] classifies thepulmonary embolism based on radiology report using CNN. L. Suliemanet.al [71] effectively classifies the patient portal messages using CNN. And for the biomedical text for recognizing the named entities in the CNN is applied in NLP platform [72].

7 Framework of NLP

In the NLP platform the commonly available framework is UIMA (Unstructured Information's Management Architecture) and GATE (General Architecture Text Engineering) which is an open-source software.

a. GATE

It was developed in 1995 at Sheffield University based on the NLP platform. it includes the basic tools of NLP for processing the low-level processes like part-speak taggers, tokenizers, penetration splitters are combined into single wrapper unit called CREOLE wrapper. In the high-level process it includes recognition of entity into ANNIE for algorithm is used in health care domain in various forms. The drawback of using machine learning algorithm is difficult to handle complex high-scale data, even though it is easy to use, simplicity, interpretability in EPHR. To overcome this drawback deep learning algorithm is used. Features of deep learning algorithm is hierarchical in construction of data features and efficient in handling long-range of data dependencies.

b. UIMA

It was initially designed by IBM since 2006 and belongs to Apache Software Foundation software. It is based on the concept of pluggable architecture and easily plug in ourcomponent in the analysis of reducing the duplication of analytical development. IBM's 2011 Jeopardy challenge Watson system has developed UIMA's framework, for recognizing the best known foundation. In addition to textual information UIMA is used to analyze the audio/video data and also it extracts the cancer features using various biomedical NLP models like MedEx,MedKAT/P, and cTAKES [73,74].

8 Challenges in EPHR using NLP

The clinical data is surveyed to focus on EPHR analyses. This information has significant features related to health care. Further this section discusses huge challengesacquired during EPHR related research.

A) Annotations Lacking

Traditional deep and machine learning techniques uses supervised model needed labelled data in the training phase. Therefore, annotating EPHR becomes challenging due to variability and cognitive complexity in providing data quality. So neural network requires training the huge amountof textual data. In some situation it only allows qualified annotator data for the training process. It is difficult to identifyit and train the model.

B) Privacy

The EPHR contains sensitive information and as per the existence law of the (US) Health based Insurance andPortability, Accountability Act (HIPAA), ensures the patient privacy in information as well as summary of treatment procedures. Therefore, before sharing of textual data in the EPHR privacy-preserving steps must be taken.

C) Interpretability

Deep neural networks produce superior results when compared it with other existing algorithms. Neural network framework requires huge parameters for training dataset which causes difficult situation of model interpretability. In the lineardata type, neural network consists of complex architecturewith non-linear layers, in which deep neural network is implemented to provide the transparency of model.

9 Datasets used in NLP

The dataset is fully available at Clinical Practice Research Data link (CPRD). Note: link, https://www.cprd.com/. For the accessibility of data https://www.cprd.com/primary-care explains: License permits full access to CPRD which has detailed terms to be used.

10 Conclusion

This paper presented the qualitative review of recent survey on the electronic health records in NLP. Also, it discussed about overview of NLP and in health care domain. Tasks involved in health care domain with NLP are discussed. Then we have presented the methods are used in the NLP for health care domain. We have provided the application areas of machine and deep learning techniques in the health care domain based on the NLP platform. Conclude this reviewpaper by explaining the challenges and difficulties in the existing techniques. Mostly available challenges are preserving privacy, interpretability and lack of annotations. This paper provides a chance to read complexity of textual information of clinical data and also it provides a challenge to the researchers for exploring the new methods in the NLP platform for health care data. In future, analysis the managing diseases, improving the quality in all aspects of the health caredomain.

References

[1]. Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, JörgKreuzer, Mark Leenay, et al. 2017. Electronic health records to facilitate clinical research. Clinical Research in Cardiology

106, 1,1–9, 2017.
[2] HIT Consultant. Why unstructured data holds the key to intelligent healthcare systems [Internet]. Atlanta (GA): HIT Consultant; cited at 2019 Jan 15, 2015.
[3]. Cao Xiao, Edward Choi, and Jimeng Sun. 2018. "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review" Journal of the American Medical Informatics Association 25, 10, 1419–1428, 2018.
[4] S. A. Hasan and O. Farri, ``Clinical natural language processing with deep learning," in Data Science for Healthcare. Springer, pp. 147-171, 2019.
[5] G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, and S. M. Meystre,``Extracting information from textual documents in the electronic healthrecord: A review of recent research," Yearbook Medical Information, vol. 17,no. 1, pp. 128-144, 2008.
[6]. Homa Alemzadeh, and Murthy Devarakonda, "An NLP-based Cognitive System for Disease Status Identification inElectronic Health Records", 978-1-5090-4179-4/17, IEEE. [7]. H. Yang, et al., "A text mining approach to the prediction of disease status from clinical discharge summaries," JAMIA,vol.16, no.4, pp.596-600, 2009.
[8]. R. Kirk and S. M. Harabagiu, "A flexible framework for deriving assertions from electronic medical records," JAMIA, vol. 18, no. 5, pp. 568-573, 2011.
[9] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, ``What can natural language processing do forclinical decision support?" J. Biomedical. Information, vol. 42, no. 5, pp. 760-772, 2009.
[10]. Essam H. Houssein, Rehab E. Mohamed, And Abdelmgeid A. Ali," Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review", Digital Object Identifier 10.1109/IEEE Access.2021.3119621, Volume 9, 2021.
[11] W. contributors, "Natural Language Processing- Wikipedia, the Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Natural_language_ processing, 2020.
[12] C. Friedman, T. C. Rindfiesch, and M. Corn, ``Natural language processing: State of the art and prospects for significant progress, a work shop sponsored by the National Library of Medicine," J. Biomed. Inform., vol. 46, no. 5, pp. 765-773, 2013.
[13] J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel, "Adoption of electronic health record systems among US non-federal acute care hospitals: 2008- 2015," ONC Data Brief, vol. 35, pp. 1-9, May 2016.
[14] L. A. Knake, M. Ahuja, E. L. McDonald, K. K. Ryckman, N. Weathers,T. Burstain, J. M. Dagle, J. C. Murray,and P. Nadkarni, ``Quality ofEHR data extractions for studies of preterm birth in a tertiary care center: Guidelines for obtaining reliable data," BMC Pediatrics, vol. 16, no. 1,p. 59, Dec. 2016.
[15] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, "Institute of medicine (US) committee on quality of health care in America," in To Err IsHuman: Building a Safer Health System. Washington, DC, USA: National Academies, 2000.
[16] J. A. Casey, B. S. Schwartz, W. F. Stewart, and N. E. Adler, ``Using electronic health records for population health research: A review ofmethods and applications," Annu. Rev. Public Health, vol. 37, no. 1, pp. 61-81, Mar. 2016.
[17] Y. Wang, ``Clinical information extraction applications: A literature review," J. Biomed. Inform., vol. 77, pp. 34-49, Jan. 2018.
[18] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones,R. Forshee, M.Walderhaug, and T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review," J. Biomed. Informat., vol. 73, pp. 14-29, Sep. 2017.
[19] L. Chen, Y. Gu, X. Ji, Z. Sun, H. Li, Y. Gao, and Y. Huang, ``Extracting medications

and associated adverse drug events using a natural language processing system combining knowledge base and deep learning," J. Amer. Med. Inform.Assoc., vol. 27, no. 1, pp. 56-64, Jan. 2020.

[20] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F.Shen, L.Wang, Y.Wang, A.Wen, Y. Zhao, S. Sohn, and H. Liu, "Clinical concept extraction: A methodology review," J. Biomed. Informat., vol. 109, Art. no. 103526, Sep. 2020.

[21] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," J. Healthcare Eng., vol. 2018, Art. no. 4302425, Apr. 2018.

[22] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe,J. B. Gutierrez, and K. Kochut, ``A brief survey of text mining: Classification, clustering and extraction techniques," 2017, arXiv:1707.02919.[Online]. Available: http://arxiv.org/abs/1707.02919, 2017.

[23] A. Finiegula, A. Poniszewska-Mara«da, and L. Chom¡tek, ``Towards the named entity recognition methods inbiomedical field," in Proc. Int. Conf. Current Trends Theory Pract. Inform. Springer, 2020, pp. 375-387, 2020.

[24] Y. Shinyama and S. Sekine, ``Proceedings of the main conference on human language technology conference of the north American chapter of the association of computationallinguistics," Assoc. Comput. Linguistics, Stroudsburg, PA, USA, Tech. Rep., 2006.

[25] B. Rink, S. Harabagiu, and K. Roberts, ``Automatic extraction of relations between medical concepts in clinical texts," J. Amer. Med. Inform. Assoc., vol. 18, no. 5, pp. 594- 600, Sep. 2011.

[26] Q.Wei, Z. Ji, Y. Si, J. Du, J.Wang, F. Tiryaki, S.Wu, C. Tao, K. Roberts, and H. Xu, ``Relation extraction from clinical narratives using pre-trained language models," in Proc. AMIA Annu. Symp., p. 1236, 2019.

[27] Y. Si and K. Roberts, ``A frame-based nlp system for cancer-relatedinformation extraction," in Proc. AMIA Annu. Symp., p. 1524, 2018.

[28] W. Sun, A. Rumshisky, and O. Uzuner, ``Evaluatingtemporal relations inclinical text: 2012 i2b2 challenge," J. Amer. Med. Inform. Assoc., vol. 20, no. 5, pp. 806-813, 2013.

[29] J. Xu, H.-J. Lee, Z. Ji, J. Wang, Q. Wei, and H. Xu," UTH_CCB system for adverse drug reaction extraction fromdrug labels at TAC-ADR 2017", 'in Proc. TAC, pp. 1-6, 2017.

[30] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki,and K. Ohe, ``Extraction of adverse drug effects from clinical records,"in Proc. MEDINFO. Amsterdam,The Netherlands: IOS Press, pp. 739-743, 2010.

[31] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VAchallenge on concepts, assertions, and relations in clinical text," J. Amer.Med. Informat. Assoc., vol. 18, no. 5, pp. 552-556, Jun. 2011.

[32] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, andM. Verhagen, ``SemEval-2016 task 12:Clinical TempEval," in Proc. 10thInt. Workshop Semantic Eval. (SemEval), 2016, pp. 1052-1062.

[33] Y. Chen, H. Cao, Q. Mei, K. Zheng, and H. Xu," Applying active learning to supervised word sense disambiguation in MEDLINE," J. Amer. Med.Inform. Assoc., vol. 20, no. 5, pp. 1001-1006, Sep. 2013.

[34] H. Liu, ``A multi-aspect comparison study of supervised word sense disambiguation," J. Amer. Med. Inform. Assoc., vol. 11, no. 4, pp. 320-331, Apr.2004.

[35] M. J. Schuemie, J. A. Kors, and B. Mons, ``Word sense disambiguationin the biomedical domain: An overview," J.Comput. Biol., vol. 12, no. 5, pp. 554-565, Jun. 2005.

[36] H. Xu, M. Markatou, R. Dimova, H. Liu, and C. Friedman, ``Machinelearning and word

sense disambiguation in the biomedical domain: Design and evaluation issues," BMC Bioinf., vol. 7, no. 1, pp. 1-16,Dec. 2006.

[37] Q. Dong and Y.Wang, ``Enhancing medical word sense inventories using word sense induction: A preliminary study," in Heterogeneous DataManagement, Polystores, and Analyticsfor Healthcare. Springer, 2020, pp. 151-16, 2020.

[38] R. M. Cronin, D. Fabbri, J. C. Denny, S. T. Rosenbloom, andG. P. Jackson, ``A comparison of rule-based and machine learningapproaches for classifying patient portal messages," Int. J. Med. Informat., vol. 105, pp. 110-120, Sep. 2017.

[39] S. Kumar Sahu and A. Anand, ``Recurrent neural network modelsfor disease name recognition using domaininvariant features,"arXiv:1606.09371. http://arxiv.org/abs/1606.09371, 2016,

[40] . M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U.Leser, ``Deep learning withword embeddings improves biomedical named entity recognition, "Bioinformatics, vol. 33, no. 14, pp. i37-i48, Jul. 2017.

[41] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, andC. Dyer, ``Neural architectures for named entity recognition,"arXiv:1603.01360. http://arxiv.org/abs/1603.01360, 2016.

[42] S. Moen and T. S. S. Ananiadou, ``Distributional semantics resources forbiomedical text processing," in Proc. LBM, pp. 39-44, 2013.

[43]. K. Raja and S. Jonnalagadda, ``Natural language processing and data mining for clinical text," Healthcare Data Anal., vol. 36, p. 219, Jan. 2015.

[44] K. P. Murphy, Machine Learning: A Probabilistic Perspective.Cambridge, MA, USA: MIT Press, 2012.

[45] O. Baclic, M. Tunis, K. Young, C. Doan, H. Swerdfeger, J. Schonfeld,P. Data, and I. Hub, ``Natural language processing (NLP) a subfield of artificial intelligence," CCDR, vol. 46, no. 6, pp. 1-10, 2020.

[46] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," J. Amer. Med. Inform. Assoc.,vol. 18, no. 5, pp. 544-551, 2011.

[47] Y. Bengio, A. Courville, and P. Vincent, ``Representation learning:A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[48] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge,MA, USA: MIT Press, 2016.

[49] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge,MA, USA: MIT Press, 2016.

[50] M. Torii, J.-W. Fan, W.-L. Yang, T. Lee, M. T. Wiley, D.
S. Zisook, andY. Huang, ``Risk factor detection for heart disease by applying textanalytics in electronic medical records," J. Biomed. Informat., vol. 58, pp. S164S170, Dec. 2015.

[51] J. C. Denny, N. N. Choma, J. F. Peterson, R. A. Miller, L.Bastarache,M. Li, and N. B. Peterson, ``Natural language processing improves identification of colorectal cancer testing in the electronic medical record,"Med. Decis. Making, vol. 32,no. 1, pp. 188197, Jan. 2012.

[52] J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, H.-J. Dai, and C.-Y. Hsu, "Identication and progression of heart disease risk factors in diabeticpatients from longitudinal electronic health records," BioMed Res. Int.,vol. 2015, pp.110, 2015.

[53] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri," Early recognitionof multiple sclerosis using natural languageprocessing of the electronichealth record," BMC Med. Informat. Decis. Making, vol. 17, no. 1, p. 24, Dec. 2017.

[54] R. L. Figueroa and C. A. Flores, ``Extracting information from electronicmedical records to identify the obesity status ofa patient based oncomorbidities and bodyweight

measures," J. Med. Syst., vol. 40, no. 8, pp. 191, Aug. 2016.
[55] S. N. Kasthurirathne, B. E. Dixon, J. Gichoya, H. Xu, Y. Xia, B. Mamlin, and S. J. Grannis, ``Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches usingplaintext medical data and non-dictionary based featureselection," J. Biomed. Informat., vol. 60, pp. 145-152, Apr. 2016.
[56] G. Napolitano, A. Marshall, P. Hamilton, and A. T. Gavin, ``Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction," Artif. Intell. Med., vol. 70, pp. 77-83, Jun. 2016.
[57] H. Yang and J. M. Garibaldi, ``A hybrid model for automatic identification of risk factors for heart disease," J. Biomed. Informat., vol. 58, pp. S171-S182, Dec. 2015.
[58] K. Buchan, M. Filannino, and Ö. Uzuner, ``Automatic prediction ofcoronary artery disease from clinical narratives," J. Biomed. Inf., vol. 72, pp. 23-32, Aug. 2017.
[59] S. M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran,T. Bekhuis, and R. S. Jacobson, ``Automated annotation and classificationof BI-RADS assessment from radiology reports," J. Biomed. Informat. vol. 69, pp. 177-187, May 2017.
[60] Q. Chen, H. Li, B. Tang, X. Wang, X. Liu, Z. Liu, S. Liu,
W. Wang,Q. Deng, S. Zhu, Y. Chen, and J.Wang, ``An automatic system to identifyheart disease risk factors in clinical texts over time," J. Biomed. Informat. vol. 58, pp.S158-S163, Dec. 2015.
[61] N.-W. Chang, H.-J. Dai, J. Jonnagaddala, C.-W. Chen, R.T.-H. Tsai,and W.-L. Hsu, ``A context-aware approach for progression tracking ofmedical concepts in electronic medical records," J. Biomed. Informat. vol. 58, pp. S150-S157, Dec. 2015.
[62] W.-W. Yim, S. W. Kwan, and M. Yetisgen, ``Classifying tumor eventattributes in radiology reports," J. Assoc. Inf. Sci. Technol., vol. 68, no. 11,p p. 2662-2674, Nov. 2017.
[63] S. N. Kasthurirathne, B. E. Dixon, J. Gichoya, H. Xu, Y. Xia, B. Mamlin,and S. J. Grannis, ``Toward better public health reporting using existingoff the shelf approaches: The value of medical dictionaries in automatedcancer detectionusing plaintext medical data," J. Biomed. Informat. vol. 69, pp.160-176, May 2017.
[64] P. L. Teixeira, W.-Q. Wei, R. M. Cronin, H. Mo, J. P. VanHouten,R. J. Carroll, E. LaRose, L. A. Bastarache, S. T. Rosenbloom,T. L. Edwards, D. M. Roden, T. A. Lasko, R. A. Dart, A. M. Nikolai,P. L. Peissig, and J. C. Denny,
``Evaluating electronic health record datasources and algorithmic approaches to identify hypertensive individuals,"J.Amer. Med. Inform. Assoc., vol. 24, no. 1, pp. 162-171, Jan. 2017.
[65] B. Tang, Z. Pan, K. Yin, and A. Khateeb, ``Recent advances of deeplearning in bioinformatics and computational biology," Frontiers Genet., vol. 10, p. 214, Mar. 2019.
[66] O. Baclic, M. Tunis, K. Young, C. Doan, and H. Swerdfeger, ``Challengesand opportunities for public health made possible by advances in naturallanguage processing," Canada Communicable Disease Rep., vol. 46,no. 6, pp. 161- 168, Jun. 2020.
[67] S. Baker, A.-L. Korhonen, and S. Pyysalo, ``Cancer hallmark text classification using convolutional neural networks," in Proc. 5th WorkshopBuilding Evaluating Resour. Biomed. Text Mining (BioTxtM), 2017, pp. 1-9.
[68] Y. Peng and Z. Lu, ``Deep learning for extracting protein-protein interactionsfrom biomedical literature,"arXiv:1706.01556. [Online]. Available: http://arxiv.org/abs/1706.01556, 2017
[69] M. Asada, M. Miwa, and Y. Sasaki, ``Extracting drug- drug interactionswith attention CNNs," in Proc. BioNLP, pp. 9-18, 2017
[70] M. C. Chen, R. L. Ball, L. Yang, N. Moradzadeh, B. E. Chapman,D. B. Larson, C. P.

Langlotz, T. J. Amrhein, and M.
P. Lungren, “Deepl earning to classify radiology free-text reports," Radiology, vol. 286, no. 3, pp. 845-852, 2017.
[71] L. Sulieman, D. Gilmore, C. French, R. M. Cronin, G. P. Jackson,M. Russell, and D. Fabbri, ``Classifying patient portalmessages usingconvolutional neural networks," J. Biomed. Informat., vol. 74, pp. 59-70, Oct. 2017.
[72] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, ``A neural networkmulti-task learning approach to biomedical named entity recognition,"BMC Bioinf., vol. 18, no. 1, p. 368, Dec. 2017.
[73]. H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, andJ. C. Denny, ``MedEx: A medication information extraction systemfor clinical narratives," J. Amer. Med. Inform. Assoc., vol. 17, no. 1, pp. 19-24, 2010.
[74] S. Doan, L. Bastarache, S. Klimkowski, J. C. Denny, and H. Xu, ``Integratingexisting natural language processing tools for medication extractionfrom discharge summaries," J. Amer.Med. Inform. Assoc., vol. 17, no. 5, pp. 528-531, Sep. 2010.