# A Survey on different semantic based machine learning techniques for Health Care data

**Majji Venkata Kishore[1], Prajna Bodapati[2]**

[1]Research Scholar, Department of CS & SE, AUCE(A) at Andhra University, Visakhapatnam, Andhra Pradesh
[2]Professor, Department of CS & SE, AUCE(A) at Andhra University, Visakhapatnam, Andhra Pradesh

Received: 13.07.2024 Revised: 14.08.2024 Accepted: 20.09.2024

**ABSTRACT**
The Electronic Health Records (EHR), the system will provide the real-time data which is used in the research. for this we use Natural Language Processing techniques , algorithms which are combined with Artificial Intelligence to extract the information. The present ERH is required to fill the potential using the NLP techniques for the direct decision of the clinical data. The Biomedical methods are combined with ML,DL, and NLP to process the ERH's and predict the information and provide the challenges and trends in the health care.

**Keywords:** Artificial Intelligence (AI), Machine Learning(ML), Natural Language Processing (NLP), Deep Learning(DL), Electronic Health Record (HER)

## INTRODUCTION

There is a better performance of using NLP in the area of medical field for the digital field by using ML technique for preprocessing the data in the research fields. Taking the clinical text and tested on the various fields for identification of different chronicle diseases. As the ML technology give the good results in the health care and by using the MLP in the analysis will be with more accuracy in the clinical test results.[1]. There is necessity of the research for the clinical test intensively to review the challenges and reviews for the information extracts[2].Using NLP in the field of the research for many use cases like the bioinformatics, computer languages and many more[1]. NLP will support the decision making for various applicators in the field of health and research, usually the unstructured data, pre-processing is necessary for the decision-support and the summarization and the tokenization, tagging, sentence splitting are included in the preprocessing to play an vital role for the information extraction[3] There are many teachings and extract in the statistical and pattern methods in ML for analyzing the clinical text and also improve the concepts and the terminologies. NLP is applied to achieve the good performance in the language modeling, The unstructured document and the identification patent data and challenges are proven with the improvement in the applications and the patient cohorts and clinical text summarization..

### A. Motivation

Extracting the clinical texts manually shows the lack of scalability and also uses for the standard data which is given by Wei et.al [4].NLP have various automatic extraction techniques and also solve with several challenges by understanding can identify technology and methods using NLP.

### B. Background

The NLP is used to process many tasks for the complex problems in different levels [5]. In many specializations The images and the information are present in the Bio-medical structure and in the recent research,[6-9]

In the chronic diseases the advance medicine will be given for the successful condition and the secondary translation application, Many researchers have done experiments by using EHR's in the application of the bio informative and health care[10-11] for the extract[2][12] modeling [13] diagnosing [14] clinical decisions. By the use of machine learning methods and deep learning methods which can understand the patient in the trajectories of the patient and can be diagnosed with the risk prediction for the given EHR's information for the development of NLP techniques in the structural form to prevent the direct decisions onset of the diseases [15]. The vast information [16] is widely recorded and remains as the written text by the descriptive methods. [17] Detection in the medical concepts [18] information n the pharamacoepidemiological free text documented records to direct the administrative processes.

**Table 1.** schema classification

| Ref.No | Schema | Codes | Example |
|---|---|---|---|
| [19] | RX Norm(Drug) | 116,075 | Morphine, Buperophie |
| [20] | CPT(Procdue) | 9,641 | Partial mastectomy, MRI Spine |
| [21] | LONIC (Labran) | 80,868 | Serum , Blood, Ethanol |

## 2. LITERUATURE

Atzal et at., [22] Identification of critical limb ischemia using NLP. NLP algorithm for PAD identification. used Dataset The Mayo clinical data warehouse. The proposed algorithm used data for a PAD cohort from a single medical center and future studies should apply and validate this algorithm to other institutions to make the findings generalizable. Atzal et at., [23] Identifying PAD cases from narrative clinical notes, NLP Algorithm. used Dataset The Mayo clinical data warehouse. A limitation of this study is that data were retrieved from the data warehouse of a single academic medical center. Leeper et at., [24] Applying Text-Mining to clinical notes for Profiling the Cilostazol Safety. NLP Algorithm, used Dataset The Stanford Translational Research Integrated Database Environment (STRIDE).A limitation on this study is that it could have missed co morbidities due to false negatives from lower sensitivity (739c). The outcome measures may not have captured events occurring outside of the hospital or that led to hospitalizations in other institutions. Buchan et al., [25] Automatic prediction of coronary artery disease from clinical narratives. Naive Bayes, MaxEnt, and SVM, used Dataset The 2014 i2b2 Heart Disease Risk Factors Challenge data set. One limitation of this experiment is embedded in the selection of patients who do not develop CAD. Another limitation of proposed study is embedded in the subsample of patients who develop CAD. The relatively small size of the dataset is a limitation of proposed study. Chen et al.,[26] Identification of heart disease risk information. Hybrid pipeline system based on both: machine learning-based rule-based approaches using: SVM, lib short Text, and CRF suite. The proposed system did neat perform very well for coronary artery disease (CAD), obesity status, and smoking status. Torri et al., [27] Detection of the risk factor for cardiac disease. A hybrid of several ML and rule-based techniques. The proposed system was not feasible to obtain objective evaluation metrics on the training set, and the current evaluation results were solely based on the one test set. Karystianis et al., [28] Extraction of the risk factors for detecting cardiac disease. Knowledge-driven and the system implement local lexicalized rules. The proposed system achieved lower performance with CAD which proved to be the most challenging class to recognize). Yang et al., [29] Extraction of the identification of heart disease risk factors. Machine learning, Rule-based methods, Dictionary-based keyword. Poorest classification accuracy is obtained for CAD The main reason for that is due to the difficulties in the identification of sentence-level CAD clinical facts, event, test, and symptom. Kogan et al., [30] Assessment of stroke severity. Several machine learning models. The current EHR database has information which could be critical for model performance including imaging of brain scans was not available. As with all studies based on real- world data, there is the potential for missing records. Healthcare information in the database was not available until January 2007. Garg et al., [31]Automatic classification of the Ischemia Stroke subtype. The proposed method relies on the level on the documentation and detail in the EHR. The system did not include the entire EHR include CT-based radiology reports to reduce variability in the dataset. Kim et al.,[32] Identifying AIS patients by automatically classifying brain MRI reports. All brain MRI Reports from a single academic intuitions. The system used text corpus which was created at a single institution. The proposed system only included brain MRI reports with conventional stroke MRI sequence. Grechishcheva et al., [33] Developed a study of risk markers identification. Supervised ML-based NLP algorithms. NLP algorithms. One of the weak sides of the algorithm is its speed. For current corpus of data, it took 6,100 seconds tea remove marginal parts of speech, short words Osborne et al., [34] Effective identification of reportable cases of cancer nationally authorized. CRCP-NLP (The Cancer Registry Control Panel) was compared only to the original manual review process, due to resource constraints. CRCP-NLP relyon proxy data, such as monthly state case submissions, to provide an estimate. Unfortunately, yearly changes in coding practices as well as multiple changes in registrar personnel and leadership make comparisons difficult. Finally, CRCP uses rules for document segmentation generated solely from UAB training data. Thus, the reliability of the used set of regular expressions 1s expected to be reduced at other institutions to the extent that provider and sectioning practices differ. AAlAbdulsalam et al.,[35] TNM stage mentions are extracted and classified from the Utah Cancer Registry records automatically. Although regular expressions were more robust for extracting TNM mentions, the used range of features used with the CRF classifier were still limited and potential improvements may be observed if other more sophisticated feature patterns were used such as character N-grams. In addition, other machine learning algorithms could yield better performance than Conditional Random Fields and further investigation is required. St Y. et al., [36 ] Frame-based NLP system based on a Bi- LSTM-CRF neural network to extract cancer-related information in clinical narratives. A limitation of this study is that its evaluation was limited to the gold standard for each process in the pipeline instead of a multi-step evaluation and optimizing the pipeline. Datta. et al., [ 37] This study presented a literature review of biomedical NLP related to cancer. Methods Natural Language Processing Machine learning. The first limitation of this review is that, given the rapid pace of NLP development .Carrell et a1., [38] Identification of breast cancer recurrences using the NLP system form clinical text. The recurrent detection system for breast cancer based on NLP. First, the used NLP modules may require adaptation to accommodate language usage and document sectioning in other institutional settings. Second, NLP development costs limit its applicability to large or repeated tasks where it is cost effective relative to 1009a manual abstraction. Third, NLP requires access to machine-readable clinical text and does not work with print documents or their scanned copies. Fourth, proposed study cohort was limited to women with early stage of charts where NLP and the reference standard were discordant.

## 3. NLP in the Health care

### A. Disambiguate in word sense

This technique will assign the sense automatically in the specific context. There are multiple terms in the clinics for interpretation including the prostate cancer for the critical issues [39-42] for the analysis in the essential [43] notes.

### B. Drug Event Detection

This technique will detect the medication such as does, allergies and reactions which are present unstructured from the clinical notes and automatically process their interactions [44-47] for the declare, laboratory results and the prescription errors.

### C. Extraction of Information

In the NLP which provides the research translation in the support of decision image translation, improvement of the quality [48] the specialization concepts of the entities which are associated for the free text .

### D. Extraction Relation

This will focuses by using the concepts in the detaches the relationships of the semantics [49-50] mentions for the disease attribute[40] identification[41] event detection[42] adverse drug events [43] extraction [41]. The clinical notes for the Integrating in the biology [45] semantic evaluation [46] clinical challenges [47] and methods for the lasts [40]

## 4. NLP methods in the biomedical fields

In the EHR, we apply the machine learning and the deep learning methods which are compared using the rule based methodology and calculate the comparisons and efficiency by the algorithms to shift from the bench marks for all algorithms [51].

### A. Rule-Based Techniques

Rule Based techniques will work on the encode pattern structure in the expressions. The rules which extend the patterns and add the relation of the negation to determine the duration. The rules can be mostly generated in 2 ways like automatically and manually in the dataset. The system can derive the efficiency and enhance the extension on adding the certain rules. Therefore the rules provide with the high precession and the rules for the data set cannot be improved by the training data set[52].

There are multiple methods based on the techniques like

1. Look up Dictionaries[53-55]
2. Ontology[56-59]
3. Set of Rules [60-62]
4. Regular Expressions Pattern[63-64]

### B. Techniques in Machine Learning

The main techniques of the Machine Learning Algorithms are classified by the Fig.

The ML algorithms like SVM, Random Forest, Logistic Regression are used for analysis and preprocessing [65]. NLP models refine through the techniques present in ML [66-67] Deep Learning Techniques: The ML take large time consumption for the feature extraction [68] in the data representation with increasing order [69] of the abstraction. The Multilayer Perceptron is a multi hidden ANN type[70] used by modern neural networks. Convolution Neural Network (CNN) is applied mainly in the domain of Image Processing by the One Dimensional series of time in the functional fitter[70]. Recurrent Neural Network (RNN) is a sequential organized data to be demonstrated in the LSTM [70].

**Table 2.** EHR Research using machine learning

| Ref.No | Model | Biomedical NLP Applications |
|---|---|---|
| [71] | SVM | Heart Disease, Diabetes, Brest Radiology |
| [72] | Naive Bayes | Heart Disease, Smoking States, Sclerosis, Obesity, Cancer |
| [73] | Random Forest | Heart Disease, Cancer, Tumor, Hypertension |

## 5. Natural Language processing in the Health Care

The NLP literature in the medical field consists of many medical terminologies. Many technologies and techniques are available to extract the exact terminology given. The goal is generally to work with the popular databases to extract the exact terminology in the literature with the outstanding scientific process.

Biology text complexity: When we talk to doctors, scientists, they specify domain-specific terminology and this communication will be in an unambiguous way. For this type of issue, we can conclude only by analyzing the terms in the field by the collections. There are three things to consider when the suggestions are given like

1. Normalization
2. Frequency
3. Statistical Analysis.

NLP is used in AI and used for all the papers used for the AI modes to push the boundaries with the limits. We need text normalization to reduce the randomness and deals to improve efficiency. When we create a large data and normalization and the text for n-grams.

Information Retrieval Normalization: The retrieve system which was built based on NLP will be present in the term expansion and wildcards. These techniques cannot deal with discrimination. The system could deal with all the specific patterns which are used to search during retrieval.

Grammar Context: A wide range of grammar will be developed based on the phenomena of linguistics. In the single form, some domains may have the highest ambiguity to give the fixed meaning. When the normalization should not be in the literal form, it should be as the constructing Schema.

The EHR are based on the conventional Static Techniques [74] logistic regression [75] support vector machine [76]

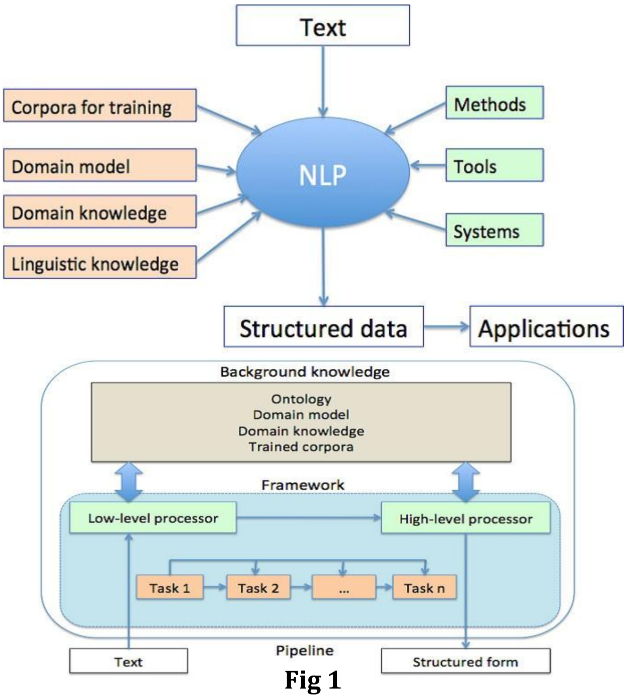

**Fig 1**

**Table 3.** Biomedical NLP Systems

| Biomedical NLP Systems | Purpose | Creator | Source of Data | Encoding |
|---|---|---|---|---|
| LSP-MLP | NLP system for extraction and summarization of signs/symptoms and drug information, and potential drugs and side effects recognition. | New York University | Progress Note. Clinical Note. X-ray Report. Discharge Summary. | SNOMED |
| MedLEE | A semantically driven system used for: i. Extracting in-formation from clinical narratives reports. ii. Participating in an automated decision-support system. iii. Allowing NLP queries | Columbia University | Radiology. Mammography. Discharge Summary. | UMLS's CUI |
| MetaMap | A highly configurable program to map biomedical text to UMLS Meta-thesaurus concepts. | National Library of Medicine | Biomedical Text Candidate and Mapping Concepts from UMLS | UMLS's CUI |
| cTAKES | Mayo clinical Text Analysis and Knowledge Extraction System. | Mayo Clinic and IBM | Discharge Summary. Clinical Note Clinical Named Entities: (diseases/disorders, signs/symptoms, anatomical sites, procedures, medications) Relation, Co-reference Smoking Status Classifier Side Effect Annotator | UMLS's CUI and RxNorm |
| SPRUS/ Sym-Text/ MPLUS | A semantically driven IE sys- tem. NLP system with syntactic and probabilistic semantic analysis | University ot Utah | Radiology Concepts from findings in radiology reports. | ICD-9 |
| SPECIALIST | A part of the UMLS project with the SPECIALIST lexi- con, semantic network, and driven by Bayesian Networks. UMLS | National Library of Medicine (NLM) | UMLS | |

## CONCLUSION

NLP using in the field of healthcare by using the techniques of the machine learning models by considering various real world use cases. Here we represented with some features by using the techniques of the clinical and the chronicle deceases. The methodologies and the challenges are associated with the open issues in the biomedical domain. The NLP will provide the best feature methods for the unstructured clinical data. Where NLP with ML provide the best models.

## REFERENCES

[1] S. A. Hasan and O. Farri, ''Clinical natural language processing with deep learning,'' in Data Science for Healthcare. Springer, 2019, pp. 147–171.

[2] G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, and S. M. Meystre, ''Extracting information from textual documents in the electronic health record: A review of recent research,'' Yearbook Med. Informat., vol. 17, no. 1, pp. 128–144, 2008.

[3] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, ''What can natural language processing do for clinical decision support?'' J. Biomed. Inform., vol. 42, no. 5, pp. 760–772, 2009.

[4] W.-Q. Wei, P. L. Teixeira, H. Mo, R. M. Cronin, J. L. Warner, and J. C. Denny, ''Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance,'' J. Amer. Med. Inform. Assoc., vol. 23, no. e1, pp. e20–e27, Apr. 2016.

[5] W. contributors. (2020). Natural Language Processing—Wikipedia, the Free Encyclopedia. Accessed: Oct. 4, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Natural_language_processing

[6] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, ''Deep learning for health informatics,'' IEEE J. Biomed. Health Inform., vol. 21, no. 1, pp. 4–21, Jan. 2017.

[7] D. Shen, G. Wu, and H. Suk, ''Deep learning in medical image analysis,'' Annu. Rev. Biomed. Eng., vol. 19, pp. 221–248, Jun. 2017.

[8] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, ''Deep learning for computational biology,'' Mol. Syst. Biol., vol. 12, no. 7, p. 878, Jul. 2016.

[9] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, ''Secondary use of EHR: Data quality issues and informatics opportunities,'' Summit Transl. Bioinf., vol. 2010, p. 1, Oct. 2010.

[10] P. B. Jensen, L. J. Jensen, and S. Brunak, ''Mining electronic health records: Towards better research applications and clinical care,'' Nature Rev. Genet., vol. 13, no. 6, pp. 395–405, 2012.

[11] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, ''A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries,'' J. Amer. Med. Inform. Assoc., vol. 18, no. 5, pp. 601–606, Apr. 2011.

[12] D. Zhao and C. Weng, ''Combining PubMed knowledge and EHR data to develop a weighted Bayesian network for pancreatic cancer prediction,''J. Biomed. Informat., vol. 44, no. 5, pp. 859–868, Oct. 2011.

[13] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, ''Using meth- ods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes,'' J. Clin. Epidemiol., vol. 66, no. 4, pp. 398–407, 2013.

[14] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, ''Natural language processing of clinical notes on chronic dis- eases: Systematic review,'' JMIR Med. Informat., vol. 7, no. 2, Apr. 2019, Art. no. e12239.

[15] K. Jensen, C. Soguero-Ruiz, K. Oyvind Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, S. Olav Skrovseth, and K. M. Augestad, ''Analysis of free text in electronic health records for identification of cancer patient trajectories,'' Sci. Rep., vol. 7, no. 1, p. 46226, May 2017.

[16] L. L. Popejoy, M. A. Khalilia, M. Popescu, C. Galambos, V. Lyons, M. Rantz, L. Hicks, and F. Stetzer, ''Quantifying care coordination using natural language processing and domain-specific ontology,'' J. Amer. Med. Inform. Assoc., vol. 22, no. e1, pp. e93–e103, Apr. 2015.

[17] H. Yang, I. Spasic, J. A. Keane, and G. Nenadic, ''A text mining approach to the prediction of disease status from clinical discharge summaries,'' J. Amer. Med. Inform. Assoc., vol. 16, no. 4, pp. 596–600, Jul. 2009.

[18] R. W. V. Flynn, T. M. Macdonald, N. Schembri, G. D. Murray, and S. F. Doney, ''Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes,'' Pharmacoepi- demiol. Drug Saf., vol. 19, no. 8, pp. 843–847, Aug. 2010.

[19] 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.

[20] H. Liu, ''A multi-aspect comparison study of supervised word sense dis- ambiguation,'' J. Amer. Med. Inform. Assoc., vol. 11, no. 4, pp. 320–331, Apr. 2004.

[21] M. J. Schuemie, J. A. Kors, and B. Mons, ''Word sense disambiguation in the biomedical domain: An overview,'' J. Comput. Biol., vol. 12, no. 5, pp. 554–565, Jun. 2005.

[22] H. Xu, M. Markatou, R. Dimova, H. Liu, and C. Friedman, ''Machine learning and word sense disambiguation in the biomedical domain: Design and evaluation issues,'' BMC Bioinf., vol. 7, no. 1, pp. 1–16, Dec. 2006.

[23] Q. Dong and Y. Wang, ''Enhancing medical word sense inventories using word sense induction: A preliminary study,'' in Heterogeneous Data Management, Polystores, and Analytics for Healthcare. Springer, 2020, pp. 151–167.

[24] 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 Is Human: Building a Safer Health System. Washington, DC, USA: National Academies, 2000.

[25] J. A. Casey, B. S. Schwartz, W. F. Stewart, and N. E. Adler, ‘‘Using electronic health records for population health research: A review of methods and applications,’’ Annu. Rev. Public Health, vol. 37, no. 1, pp. 61–81, Mar. 2016.

[26] Y. Wang, ‘‘Clinical information extraction applications: A literature review,’’ J. Biomed. Inform., vol. 77, pp. 34–49, Jan. 2018.

[27] 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 informa- tion: A systematic review,’’ J. Biomed. Informat., vol. 73, pp. 14–29, Sep. 2017.

[28] 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 lan- guage processing system combining knowledge base and deep learning,’’J. Amer. Med. Inform. Assoc., vol. 27, no. 1, pp. 56–64, Jan. 2020.

[29] 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 extrac- tion: A methodology review,’’ J. Biomed. Informat., vol. 109, Sep. 2020, Art. no. 103526.

[30] Y. Shinyama and S. Sekine, ‘‘Proceedings of the main conference on human language technology conference of the north American chap- ter of the association of computational linguistics,’’ Assoc. Comput. Linguistics, Stroudsburg, PA, USA, Tech. Rep., 2006.

[31] 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.

[32] 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., 2019, p. 1236.

[33] Y. Si and K. Roberts, ‘‘A frame-based nlp system for cancer-related information extraction,’’ in Proc. AMIA Annu. Symp., 2018, p. 1524.

[34] W. Sun, A. Rumshisky, and O. Uzuner, ‘‘Evaluating temporal relations in clinical text: 2012 i2b2 challenge,’’ J. Amer. Med. Inform. Assoc., vol. 20, no. 5, pp. 806–813, 2013.

[35] J. Xu, H.-J. Lee, Z. Ji, J. Wang, Q. Wei, and H. Xu, ‘‘UTH_CCB system for adverse drug reaction extraction from drug labels at TAC-ADR 2017,’’ in Proc. TAC, 2017, pp. 1–6.

[36] 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, 2010, pp. 739–743.

[37] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, ‘‘2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text,’’ J. Amer. Med. Informat. Assoc., vol. 18, no. 5, pp. 552–556, Jun. 2011.

[38] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. Verhagen, ‘‘SemEval-2016 task 12: Clinical TempEval,’’ in Proc. 10th Int. Workshop Semantic Eval. (SemEval), 2016, pp. 1052–1062.

[39] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, ‘‘2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records,’’ J. Amer. Med. Inform. Assoc., vol. 27, no. 1, pp. 3–12, Jan. 2020.

[40] R. M. Cronin, D. Fabbri, J. C. Denny, S. T. Rosenbloom, and G. P. Jackson, ‘‘A comparison of rule-based and machine learning approaches for classifying patient portal messages,’’ Int. J. Med. Infor- mat., vol. 105, pp. 110–120, Sep. 2017.

[41] K. Raja and S. Jonnalagadda, ‘‘Natural language processing and data min- ing for clinical text,’’ Healthcare Data Anal., vol. 36, p. 219, Jan. 2015.

[42] Y. Ni, J. Wright, J. Perentesis, T. Lingren, L. Deleger, M. Kaiser, Kohane, and I. Solti, ‘‘Increasing the efficiency of trial-patient match- ing: Automated clinical trial eligibility pre-screening for pediatric oncol- ogy patients,’’ BMC Med. Informat. Decis. Making, vol. 15, no. 1, p. 28, Dec. 2015.

[43] M. Small, D. H. Kiss, Y. Zlatsin, D. L. Birtwell, H. Williams, M. A. Guerraty, Y. Han, S. Anwaruddin, J. H. Holmes, J. A. Chirinos, R. L. Wilensky, J. Giri, and D. J. Rader, ‘‘Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease,’’ J. Biomed. Informat., vol. 72, pp. 77–84, Aug. 2017.

[44] Y. Lu, C. J. Vitale, P. L. Mar, F. Chang, N. Dhopeshwarkar, R. A. Rocha, and L. Zhou, ‘‘Representation of information about family relatives as structured data in electronic health records,’’ Appl. Clin. Informat., vol. 5, no. 2, pp. 349–367, 2014.

[45] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, ‘‘Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,’’ Sci. Rep., vol. 6, no. 1, pp. 1–10, May 2016.

[46] N. Afzal, V. P. Mallipeddi, S. Sohn, H. Liu, R. Chaudhry, C. G. Scott, J. Kullo, and A. M. Arruda-Olson, ''Natural language processing of clinical notes for identification of critical limb ischemia,'' Int. J. Med. Informat., vol. 111, pp. 83–89, Mar. 2018.

[47] N. J. Leeper, A. Bauer-Mehren, S. V. Iyer, P. LePendu, C. Olson, and N. H. Shah, ''Practice-based evidence: Profiling the safety of cilostazol by text-mining of clinical notes,'' PLoS ONE, vol. 8, no. 5, May 2013, Art. no. e63499.

[48] R. Vijayakrishnan, S. R. Steinhubl, K. Ng, J. Sun, R. J. Byrd, Z. Daar, B. A. Williams, C. deFilippi, S. Ebadollahi, and W. F. Stewart, ''Preva- lence of heart failure signs and symptoms in a large primary care popu- lation identified through the use of text and data mining of the electronic health record,'' J. Cardiac Failure, vol. 20, no. 7, pp. 459–464, Jul. 2014.

[49] Z. Tian, S. Sun, T. Eguale, and C. M. Rochefort, ''Automated extraction of VTE events from narrative radiology reports in electronic health records: A validation study,'' Med. Care, vol. 55, no. 10, p. e73, 2017.

[50] B. E. Chapman, S. Lee, H. P. Kang, and W. W. Chapman, ''Document- level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm,'' J. Biomed. Informat., vol. 44, no. 5, pp. 728–737, Oct. 2011.

[51] F. P.-Y. Lin, A. Pokorny, C. Teng, and R. J. Epstein, ''TEPAPA: A novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records,'' Sci. Rep., vol. 7, no. 1, pp. 1–13, Dec. 2017.

[52] C. Nath, M. S. Albaghdadi, and S. R. Jonnalagadda, ''A natural language processing tool for large-scale data extraction from echocardiography reports,'' PLoS ONE, vol. 11, no. 4, Apr. 2016, Art. no. e0153749.

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

[54] 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.

[55] 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.

[56] 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.

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

[58] C. Friedman, T. C. Rindflesch, and M. Corn, ''Natural language process- ing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine,'' J. Biomed. Inform., vol. 46, no. 5, pp. 765–773, 2013.

[59] 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.

[60] M. Small, D. H. Kiss, Y. Zlatsin, D. L. Birtwell, H. Williams, M. A. Guerraty, Y. Han, S. Anwaruddin, J. H. Holmes, J. A. Chirinos, R. L. Wilensky, J. Giri, and D. J. Rader, "Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease," J. Biomed. Informat., vo1. 72, pp. 77—84, Aug. 2017.

[61] Y. Lu, C. J. Vitale, P. L. Mar, F. Chang, N. Dhopeshwarkar, R. A. Rocha, and L. Zhou, "Representation o1 information about family relatives as structured data in electronic health records," Appl. Clin. Informat., vol. 5, no. 2, pp. 349—367, 2014.

[62] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," Sci. Rep., vol. 6, no. 1, pp. 1—10, May 2016.

[63] N. Atzal, V. P. Mallipeddi, S. Sohn, H. Liu, R. Chaudhry, C. G. Scott, J. Kullo, and A. M. Arruda-Olson, "Natural language processing of clinical notes for identification of critical limb ischemia," Ant. J. Med. Informat., vol. 111, pp. 83—89, Mar. 2018.

[64] N. J. Leeper, A. Bauer-Mehren, S. V. lyer, P. LePendu, C. Olson, and N. H. Shah, "Practice-based evidence: Profiling the safety of cilostazol by text-mining of clinical notes," PLoS ONE, vol. 8, no. 5, May 2013, Art. no. e63499.

[65] R. Vijayakrishnan, S. R. Steinhubl, K. Ng, J. Sun, R. J. Byrd, Z. Daar, B. A. Williams, C. deFilippi, S. Ebadollahi, and W. F. Stewart, "Preva- lence of heart failure signs and symptoms in a large primary care popu- lation identified through the use of text and data mining o1 the electronic health record," 7. Cardiac Failure, vo1. 20, no. 7, pp. 459—464, Jul. 2014.

[66] Z. Tian, S. Sun, T. Eguale, and C. M. Rochefort, "Automated extraction of VTE events trom narrative radiology reports in electronic health records: A validation study," Med. Care, vo1. 55, no. 10, p. e73, 2017.

[67] B. E. Chapman, S. Lee, H. P. Kang, and W. W. Chapman, “Document- level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm,” J. Biomed. Informat., vo1. 44, no. 5, pp. 728-737, Oct. 2011.

[68] F. P.-Y. Lin, A. Pokorny, C. Teng, and R. J. Epstein, “TEPAPA: A novel in silico feature learning pipeJ ine for mining prognostic and associative factors from text-based electronic medical records,” S‹'i. key., vol. 7, no. 1, pp. 1—13, Dec. 2017.

[69] C. Nath, M. S. Albaghdadi, and S. R. Jonnalagadda, “A natural language processing tool for large-scale data extraction from echocardiography epoRs,” PLoS ONE, vol. 11, no. 4, Apr. 2016, Art. no. e0153749. K. P. Murphy, Machine Learning.’ A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[70] 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.

[71] 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.

[72] N. J. Nilsson and N. J. Nilsson, Artifie'ial Intelligence. A New Synthesis. San Mateo, CA, USA: Morgan Kaufmann, 1998.

[73] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mac'h. Intell., vol. 35, no. 8, pp. 1798—1828, Aug. 2013.

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

[75] Y. Goldberg, “A primer on neural network models for natural language processing,” J. Artif. Intell. Res., vol. 57, pp. 345—420, Nov. 2016.

[76] P. Domingos, “A few useful things to know about machine learning,” Commun. ACM, vol. 55, no. 10, pp. 78—87, 2012.

[77] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in Proc. 25th Int. Conj. Mach. Learn. (ICML), 2008, pp. 1096—1103.

[78] F. E. Harrell, Jr., K. L. Lee, R. M. Califl, D. B. Pryor, and R. A. Rosati, “Regression modelling strategies for improved prognostic prediction,” Statist. Med., vol. 3, no. 2, pp. 143—152, Apr. 1984.