International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

The Interplay between Natural Language Processing (NLP) and Clinical Data Mining in Healthcare: A Review

Shashank Agarwal*[1], Praveen Gujar, Sriram Panyam[3]

Submitted: 25/01/2024 **Revised**: 03/03/2024 **Accepted**: 11/03/2024

Abstract: Natural Language Processing (NLP) has evolved as a transformational force in the healthcare industry, which suggests innovative ways to extract, generate, and process clinical data. This review paper delves into the critical role of NLP in recasting the healthcare sector through the extraction of essential medical information from multiple sources, such as Electronic Health Records (EHRs). The objective of this review is to identify the crucial role of NLP in the healthcare industry by extracting medical information from clinical data thus augmenting patient care, disease detection, clinical decision-making, patient compliance, and even medical transcription. This review includes a detailed assessment of various NLP techniques, from rule-based techniques to statistical approaches and large language models using transfer learning. It then explores different NLP libraries and frameworks that are being widely used in a variety of fields. It also covers the types of clinical data that can be further refined and utilized through NLP, followed by several most common NLP libraries that are particularly adapted to each healthcare application. Moreover, applications and uses of NLP in healthcare systems are also discussed, paving the way for further research and its future scope in the field of health information technology. Although NLP holds a strong promising position in patient care, however, linguistic diversity, unstructured data, and semantic ambiguities are among some of the significant challenges and barriers to its implementation. Therefore, the article aims to highlight the necessity for continuous improvement and advancement in NLP techniques for ensuring the accuracy, efficiency, and reliability of data extraction, interpretation, and application within the healthcare domain.

Keywords: *natural language processing, healthcare, clinical data, electronic health records, transfer learning, python library*

1. Introduction

Natural Language Processing (NLP) has emerged as a field of research and application that examines how computers may be utilized to comprehend and alter natural language speech and text to give beneficial outcomes [11]. This text can be of various types, most commonly in the form of electronic health records (EHRs). EHRs are prepared on a routine basis and can provide supplementary information regarding the clinical background of the patient and the effectiveness of treatments in real-world scenarios [7]. Nowadays, a large number of data is being used in numerous areas of scientific research because it facilitates the investigation of complex problems measured at different times [53]. This technique of data interpretation is also common in medical sciences, particularly with the introduction of algorithms that have the capability to customize treatments for various disorders. Such techniques for treatment aim at improving accuracies in both prognosis and diagnosis. EHRs cover a broad variety of longitudinal data, ranging from medicine prescriptions to environmental variables [35]. This enables the assessment of complex interactions between the effects of treatment and auxiliary variables, including information on various symptoms or several previous episodes.

Previous research has manifested the utility of employing NLP models when extracting medical information from EHRs [20]. However, to gain an acceptable accuracy and flexibility level, NLP models also need input from the trained medical staff in annotating medical documentation. Statistical learning models assume relations from data but need a certain degree of data preparation and structuring in order for its implementation [42]. This kind of structuring is carried out by adding supplemental data to medical texts, where physicians highlight lengths of text that describe medical concepts of interest. The level of input required is sufficiently low in contrast with the manual reviews, while the use of the “state-of-the-art” NLP procedures further decreases the need for an extensive annotation process. Direct participation is mandatory from medical practitioners when using EHRs. This direct participation by medical professionals, as data generators (i.e., for generating coded text), is important in determining the relevancy of extracted medical information from the EHRs [46].

The review paper aims to highlight the crucial role of NLP in the healthcare industry by extracting medical information from clinical data thus augmenting patient care, disease detection, clinical decision-making, patient compliance, and even medical transcription.

[1] *Independent Researcher, Chicago, IL, USA*
ORCID ID: 0009-0003-7679-6690
[2] *Director – LinkedIn, San Francisco, CA, USA*
ORCID ID: 0009-0008-9905-1751
[3] *CTO – DagKnows, San Francisco, CA, USA*
ORCID ID : 0009-0006-0025-9110
** Corresponding Author Email: shashanka757@gmail.com*

2. Methodology

A comprehensive search of the available research literature was conducted using the databases PubMed, MEDLINE, and Google Scholar. The key terms used were “Natural language processing”, “medical data”, “clinical text”, “health care industry”. 42 articles were selected from the vast literature present in database, that followed the inclusion criteria. This criteria for the inclusion of the sources included whether they have been peer-reviewed, whether they are published in English and whether they are relevant to the topic under study. Sources that had not been evaluated by other researchers, articles that were not immediately relevant to the topic under investigation, and sources written in languages apart from English were all excluded from the review. Specifically, the articles that focused on NLP application in some other sector, rather than healthcare, were also excluded from the article.

3. Various Techniques in NLP

There are three major stages occurring in an NLP pipeline for data extraction (as shown in Figure 1), i.e., preprocessing of raw text, feature extraction, and Artificial Intelligence (AI) modeling. In the first stage, NLP takes text or speech in the form of raw data input and preprocesses it. The second stage involves the numerical feature extraction of the clean data which was initially obtained after preprocessing. In the final stage, AI models are finally built, having the extracted numerical features to produce the output data, for performing particular NLP tasks.

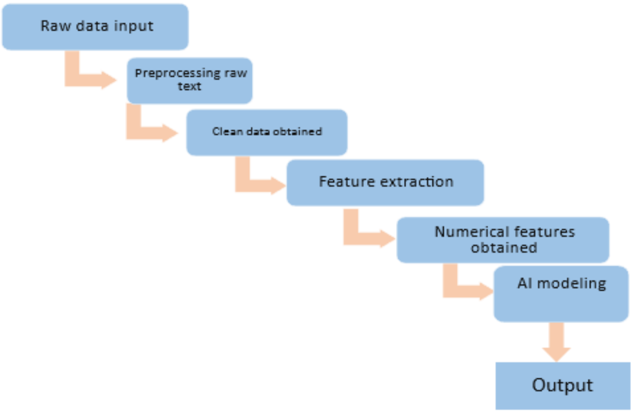

Fig. 1. The Stages of NLP Pipeline.

3.1. Rule Based Techniques

Grammar serves as a key component of numerous applications for natural language processing (NLP), such as rule-based machine translation. Grammar formally explains the language structure and the way in which different words are collected to form language sentences. NLP grammars are different from each other in their coverage, grammar formalisms that are used, and the linguistic postulations on which they are based. Moreover, when the focus is to obtain more reliable and expected results instead of robustness, rule-based language processing is mostly preferred such as in recognition of speech [44]. Different methods to reduce the burden of grammar development have been examined. These techniques include “domain-specific grammar development”, “grammar adaptation”, and “grammar sharing”. Domain-specific grammar development is used for grammar that covers a domain-specific language or a sublanguage, while grammar adaptation is useful in generating grammar for a new language by reusing the data from an existing grammar of a language. In grammar sharing, the rules of grammar are shared among various languages, instead of only reusing the data of an existing grammar. The grammar-sharing technique has only been applied to languages that are closely related [8]. A better coherent analysis, reusing of code, decrease in the number of rules, and ease of modifying and debugging on one grammar instead of many are some of the salient advantages of grammar sharing technique.

3.2. Statistical Techniques

In statistical methods, natural languages are processed by frequency and probability. A training corpus is used for the calculation of probability. A number of techniques can be utilized to apply statistical methods, such as Hidden Markov Model (HMM), N-gram, Likelihood Estimation, Decision Trees, Conditional Random Fields, Maximum Entropy, and Support Vector Machines [29]. This shows that simple distributional methods are beneficial for solving complex engineering problems that seemed almost resistant to the implementation of a priori knowledge [32].

HMM is recognized as one of the distinguished probabilistic models, employed for working on different language processing problems. This model performs by assigning the joint probability to paired observation as well as the label sequence. Later, the parameters are then trained to enhance the joint probability of training datasets [30]. Maximum Entropy Markov Model or MEMM is another conditional probabilistic sequence model, representing multiple characteristics of a word and handling long-term dependency. Based on the concept of maximum entropy, it follows that the model which is least biased and evaluates all known facts is the one that is responsible for maximizing entropy [49]. Conditional Random Fields is comparatively a new mathematical and probabilistic model that may be used to solve problems involving sequence labeling and tagging for the input data [48]. CRFs, also known as random fields, are basically undirected graphical models used for calculating the conditional probability of potential output nodes that correspond to the input. Yet again, another model known as the Support vector machines or SVM, are primarily suitable for text categorization and solving problems that involve two-class pattern recognition in NLP [5].

In N-GRAM statistical technique, a word is provided with a tag based on the “probability of that tag to occur with that word’. This probability is then calculated afterward from a pre-annotated corpus and used to provide the accurate tag to the word present in the tested corpus. Last but not least, Genetic Algorithms are also capable of solving a wide range of problems with the aim to search for a better solution among many available solutions [47], Thus, statistical approaches play a critical role in language processing and almost all the NLP applications are able to be developed by either employing one or through the combination of more than one statistical method.

3.3. Large Language Model Employing Transfer Learning

Deep learning models most often require large amounts of data. Nevertheless, these huge datasets are not always achievable. This is very much common in numerous challenging NLP works. Moreover, deep learning models also require huge computing resources. These limitations motivate researchers to work on the possibility of knowledge transfer utilizing large trained models. Due to the emergence of large models, the need for transfer learning is increasing day by day. The basic concept behind transfer learning is the transfer of knowledge from one trained model to the other model. Depending on the availability of labeled datasets, the process of transfer learning is classified into transductive learning and inductive learning [4].

Transfer learning is general and might be applied to a number of tasks involving machine learning. In recent times, transfer learning has been applied greatly in NLP which has also been explained by [28], discussing the evolution of transfer learning particularly in NLP. They majorly focused on the most paramount transfer learning approach which is known to be sequential fine tuning. Still, however, transfer learning in NLP must be studied more deeply with a focus on all transfer learning approaches.

4. Libraries and Frameworks for NLP

This section of the review describes widely employed NLP software libraries and frameworks for research and application development. It is worth noting that these tools are continuously in the process of evolving and it is not suitable to call one framework than the other, since their appropriateness is dependent on the application domain.

Natural Language Toolkit or NLTK is an open-source, vastly employed Python library for natural language processing [38]. Various algorithms are available for the tokenization of text, stemming, parsing, classification, removal of stop words, PoS tagging, clustering, and semantic reasoning. NLTK also provides wrappers for other various libraries of NLP. A prominent feature of NLTK is that it gives access to more than 50 lexical and corpora resources e.g., the WordNet.

Stanford CoreNLP is yet another open-source, integrated Java toolset for research and application development of natural language processing [51]. It is extensible and is capable of running as an easy-to-access simple web service. It is also multilingual and thus encompasses a variety of languages including Arabic, French, Chinese, Spanish, English, and German. It also gives support for the annotation of arbitrary texts and combines some of Stanford’s NLP tools, such as parsers, PoS tagger, coreference resolution system, named entity recognizer, bootstrapped pattern learning, sentiment analyzer, and open information extraction methods. The Apache OpenNLP is very similar to the Stanford CoreNLP Toolset in functionality and easy use. However, they both differ in the licensing terms. Due to its active maintenance and evolution, Stanford CoreNLP has a slight edge over Apache OpenNLP.

GATE, i.e., General Architecture for Text Engineering is also a Java toolset for natural language processing [52]. It is also an integrated form that was developed at the University of Sheffield in 1995, and since then it has been under continuous development. GATE has a wide range of functions for analytics and processing of text. It can also be employed for generating and defining text-processing workflows.

MALLET i.e., Machine Learning for LanguagE Toolkit is known as a library of Java software for statistical Natural language processing [33]. This library was developed and is still maintained at the University of Massachusetts (Amherst). MALLET features different algorithms for the classification of documents, sequence tagging, etc. For instance, it gives an effective implementation of various algorithms such as Pachinko Allocation, Latent Dirichlet Allocation (LDA), and Hierarchical LDA.

OpenNLP and CoreNLP are tools for social media NLP libraries that perform well on documents that are formally written, while not so well on documents that are short and informal. TwitterNLP is a very good example of a Python library for carrying out NLP pipeline operations on small documents e.g., tweets. Particularly, its primary features include chinking (grouping of words and transforming them into meaningful phrases by the use of PoS tags), tokenization, named entity recognition, and PoS tagging. Employed in the identification of tweet contents, tokenization, URLs, PoS tagging and emoticons, TweetNLP is also a Java library [19].

5. Types of Clinical Data

Clinical data covers a broad range of information that is gathered during medical research as well as patient care. Below are described some of the major types of clinical data that are frequently used in healthcare:

5.1. Electronic Health Records or EHRs

These digital records contain detailed patient information, which includes the patient's medical history, disease diagnoses, any current or previous medications, treatment plans, and results from laboratory tests. These records are crucial for patient care, research in the clinical domain, and healthcare administration [2].

5.2. Medical Imaging Data

This type of clinical data includes CT scans, X-rays, MRIs, and other different medical images that are routinely used for disease diagnosis and strategic treatment planning.

5.3. Laboratory Data

Laboratory data consists of the findings and results obtained from urine tests, blood tests, genetic tests, and other various diagnostic procedures. Laboratory information provides insights into the status of patient's health and potential diseases [34].

5.4. Clinical Trials Data

This type of data is obtained from clinical trials and includes information regarding participants of the study, treatment protocols, outcomes of research, and adverse events of medications or treatment from clinical research studies. Clinical trials data is essential in drug development as well as evidence-based medicine.

5.5. Patient-Reported Outcomes (PROs)

This data helps in recording patient-reported symptoms of the disease, side-effects and adverse effects of the medication/ treatment, drug-drug or drug-food interactions, quality of life, and satisfaction of treatment. PROs help in assessing the effect of healthcare interventions done on patients' health and well-being [6].

5.6. Genomic Data

This type of data includes information regarding a patient's DNA, RNA, and other genetic variations in DNA sequencing. Genomic data is very crucial in healthcare comprehending genetic factors in various diseases and thus customization of medications [31].

5.7. Vital Signs Data

These include measurements obtained from body vitals for example body temperature, blood pressure, respiratory rate, and heart rate or pulse. These measurements give immediate insights into the physiological state of the patient [50].

5.8. Medical Billing Data

Medical billing data comprises information regarding healthcare services that are provided to either inpatients or outpatients, costs, and insurance claims and reimbursements. This type of data is used for billing, reimbursement procedures, and analysis of healthcare utilization [21].

6. NLP Libraries Used in the Healthcare Sector

6.1. spaCy

spaCy is an open-source Python NLP library that is widely known for its features like fast speed and functional efficiency. It is specifically built to process medical texts for recognizing and classifying named entities e.g., medical terminologies, drugs, anatomical parts, various diseases, methods and procedures, etc. [15].

6.2. scispaCy

scispaCy is another Python library used for processing biomedical and clinical text. It is built on the spaCy model, being its specialized version, and comprises pre-trained models for different biomedical tasks namely dependency parsing, entity recognition, and tagging part of speech [37].

6.3. BioBERT

BioBERT is yet another advanced, trained model representing contextual language especially developed for scientific and clinical text mining and processing. Its concept is based on the architecture of BERT architecture and is priorly trained with PubMed, Wiki, PMC, and Books. BioBert is fine-tuned on clinical data, making it appropriate for different NLP tasks in the healthcare sector (Lee et al., 2020).

6.4. ClinicalBERT

ClinicalBERT is another model based on BERT and adjusted particularly for processing clinical notes as well as discharge summaries. It is developed for capturing clinical semantics and is beneficial for tasks e.g., clinical text classification, NER or named entity recognition, etc. It is used to process the data obtained from MIMIC-3 databases [23].

6.5. Dismod-ML

This probabilistic model is a machine learning framework utilized for modeling and predicting disease burden, risk factors for a particular disease, etc. It makes use of NLP techniques for the extraction of data and its analysis in the domain of global health [17].

6.6. MedNLP

MedNLP is a clinical Python library that focuses on performing various clinical NLP tasks such as named entity recognition (NER), questioning and answering, and extraction of relationships between various nodes and variables from the clinical text [25].

7. Uses and Applications of NLP in Healthcare

7.1. Clinical Decision Support System or CDSS

Clinical Decision support systems or CDSS help healthcare practitioners in making various clinical decisions related to the best treatment option for the patient, adjusting risk factors, etc. [12]. The characteristics of the patient are compared to the CDSS database to produce recommendations that might be considered by physicians during medical decision-making [36]. CDSS can search patient history, current patient condition, and other parameters. Treatment options will then be described by CDSS for the patient, depending on a standard guideline. There is also a powerful connection between CDSS and customized medication [14]. For instance, the framework that was constructed for oncology by [9] addresses the complex management of cancer patients and combines this knowledge for offering personalized medication.

7.2. Patient Care, Disease Detection, and Adverse Drug Events Detection

NLP can be utilized in processing EHRs for extracting important information including drugs, disease diagnosis, symptoms of diseases, etc. This data in turn can be used in improving patient care and patient compliance. Moreover, massive amounts of data related to drug safety can also be processed for the identification of adverse drug reactions (ADRs) and drug interactions. Any outbreak of infectious epidemic or diseases can also be identified by processing scientific articles, news, or media posts [27].

7.3. Image Captioning

Every day, clinicians need to examine numerous medical images and write medical reports. Using NLP, medical image captioning can be made for generating textual descriptions of any given medical photograph such as a chest X-ray, which will aid in speeding up the process along with saving precious time and effort for practitioners [40].

7.4. Speech-to-Text Transcription for Electronic Health Records (EHRs)

EHRs act as a warehouse of all the information related to patients' health. Using an EHR of high quality in the healthcare sector is important to prevent medical errors which could be achieved by the integration of NLP in generating EHRs., Electronic health records now provide a different method of using keyboards and templates for generating medical progress notes which are based on the voice of a physician. This system was assessed in a randomized clinical trial to check whether or not the notes created through voice resulted in improved timeliness of availability and quality of notes [41].

7.5. Speech-to-Text Transcription for Electronic Prescribing

The process of digitization of medical prescriptions results in the provision of high accuracy and reliability. In a study, the researchers tried to develop a “smartphone-based” electronic prescribing system utilizing NLP for reducing the cost it takes the physician's voice and transcribes it into medical text, and later on print an E-prescribing or can even save it as pdf file in mobile phones [26].

7.6. Medical Robots and Chatbots

Robots are being used in the healthcare sector for the rehabilitation of patients such as exoskeletons and Robotic Endoscopy. As NLP is able to represent the comprehension of text and speech, it may act as a sensing model for the interaction of humans and robots. In a study conducted by [54], system was proposed for the detection and understanding of a surgeon’s natural language, then afterward translating it into robotic-executive commands utilizing speech recognition methods and extraction. Chatbots, also called conversational agents, are one of the major technical solutions to the problem of lack of mental health workforce. For instance, the chatbot “Wysa” makes use of certain evidence-based therapies like cognitive therapy, in order to target depression symptoms for its users. Another chatbot i.e., LISSA provides training for autism patients to develop social skills in them [3]. SPeCECA is yet another chatbot [53], assisting victims or incident witnesses. Thus, a person not having any first aid skills may help the victim in surviving by applying first aid support by following virtual assistance [1].

Table 1 provides a more detailed understanding of the NLP applications and the type of techniques used in performing a number of different healthcare tasks and functions.

Table 1. Description of NLP applications in various healthcare functions

Category	Applications	Related NLP Approaches
Collection of data and clinical communication	Communication between patient and provider [43], clinical documentation [18]	Machine translation and recognition of speech
Data management	Eases the process of data retrieval, questioning/ answering [24], and managing clinical documents [56]	Text generation and information extraction
Screening of population	Analysis of healthcare data in questionnaires and surveys [16]	Information extraction
Allocation of medical resources	Forecasting and reduction in patient readmission rates, virtual assistants, robots, prioritization	Speech recognition, information extraction

	of patient treatment (patient triage) [22]	
Quality control of services	Improving service quality and experience of patients [13]	Question answering and information retrieval
Clinical decision support	Provides evidence for clinical decisions by building CDSS [57]	Information extraction
Personal health assistants	Enabling online, remote healthcare services	Speech recognition and information retrieval
Clinical and pre-clinical research	Designing and analysis of clinical trials, screening of drugs	
Medical education and awareness	Ease in obtaining, representing, and access of medical knowledge [45]	Machine translation, question/answering, knowledge engineering
Drug discovery	Designing of new molecules [39]	Knowledge engineering, extraction, and retrieval of information
Review of drugs and safety monitoring	Monitoring of adverse drug events [55]	Information extraction

8. Barriers to Integrating NLP into Healthcare

For achieving widespread applications of NLP in the healthcare sector, it is needed that existing NLP models be adapted in order to work in various settings under different situations. This adoption process is extremely challenging due to the fact that it is dependent on the linguistic content, and structure, along with other features of the current NLP system which will be heterogeneous with the advanced system [10]. Other challenges subjected to NLP itself are also considerable such as the language barrier because many languages exist in the world with specific grammar and vocabulary so it will take immense effort to process text which is in a language other than English language. Moreover, apart from the challenges in extracting accurate meaning from the text, various accents of a single language, unstructured data, multipurpose words, etc., the most complex challenge regarding linguistics is known as Crash Blossom i.e., syntactic unclarity in languages. Therefore, these challenges and limitations show there is a need for further refinement of NLP applications in this domain of research.

9. Conclusion

Natural Language Processing (NLP) is continuously emerging as a field of research and application exploring the way computer technology may impact humans by altering natural language speech and text. Research shows that NLP also has numerous applications in the medical field in mining and processing clinical data such as when extracting medical information from EHRs. The emergence of Transfer learning models and large language models offer possible solutions for data insufficiency, while different NLP software libraries provide necessary tools for the researchers in health domain. However, integration of NLP into healthcare also faces several hurdles such as language diversity, the unstructured nature of clinical data, various accents, etc. points out the need for continuous advancements in NLP approaches for ensuring accuracy and relevance in data interpretation and application.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] M. Abdelwahap, M. Elfarash, and A. Eltanboly, "Applications of Natural Language Processing in Healthcare Systems," in The International Undergraduate Research Conference, vol. 5, no. 5, pp. 111-115, Aug. 2021.

[2] J. Adler-Milstein et al., "Electronic health record adoption in US hospitals: progress continues, but challenges persist," Health Affairs, vol. 34, no. 12, pp. 2174-2180, 2015.

[3] A. Ahmed et al., "Anxiety and depression chatbot features: a scoping review," JMIR Preprints, 26341, 2020.

[4] Z. Alyafeai, M. S. AlShaibani, and I. Ahmad, "A survey on transfer learning in natural language processing," arXiv preprint arXiv:2007.04239, 2020.

[5] P. J. Antony and K. P. Soman, "Kernel based part of speech tagger for kannada," in 2010 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2139-2144, Jul. 2010.

[6] E. Basch et al., "Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial," Journal of Clinical Oncology, vol. 34, no. 6, p. 557, 2016.

[7] C. Bombardier and A. Maetzel, "Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies?," Annals of the Rheumatic Diseases, vol. 58, suppl. 1, pp. I82-I85, 1999.

[8] P. Bouillon et al., "A shared multilingual grammar for machine speech translation," in Proceedings of the

13th Conference on Natural Language Processing. Long Articles, pp. 93-102, Apr. 2006.

[9] A. Bucur et al., "Workflow-driven clinical decision support for personalized oncology," BMC Medical Informatics and Decision Making, vol. 16, pp. 151-162, 2016.

[10] D. S. Carrell et al., "Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings," Journal of the American Medical Informatics Association, vol. 24, no. 5, pp. 986-991, 2017.

[11] K. Chowdhary and K. R. Chowdhary, "Natural language processing," in Fundamentals of Artificial Intelligence, pp. 603-649, 2020.

[12] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?," Journal of Biomedical Informatics, vol. 42, no. 5, pp. 760-772, 2009.

[13] K. Doing-Harris, D. L. Mowery, C. Daniels, W. W. Chapman, and M. Conway, "Understanding patient satisfaction with received healthcare services: a natural language processing approach," in AMIA Annual Symposium Proceedings, vol. 2016, p. 524, 2016.

[14] G. J. Downing, S. N. Boyle, K. M. Brinner, and J. A. Osheroff, "Information management to enable personalized medicine: stakeholder roles in building clinical decision support," BMC Medical Informatics and Decision Making, vol. 9, pp. 1-11, 2009.

[15] Explosion AI, "spaCy: Industrial-strength natural language processing," online, Available: https://spacy.io, 2018.

[16] D. Georgiou, A. MacFarlane, and T. Russell-Rose, "Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools," in 2015 Science and Information Conference (SAI), pp. 352-361, Jul. 2015.

[17] C. Gopalappa et al., "Dismod-ML: A Python framework for disease modeling," PLoS ONE, vol. 14, no. 6, p. e0217976, 2019.

[18] F. R. Goss et al., "A clinician survey of using speech recognition for clinical documentation in the electronic health record," International Journal of Medical Informatics, vol. 130, p. 103938, 2019.

[19] V. N. Gudivada and K. Arbabifard, "Open-source libraries, application frameworks, and workflow systems for NLP," in Handbook of Statistics, vol. 38, pp. 31-50, 2018.

[20] K. Haerian, H. Salmasian, and C. Friedman, "Methods for identifying suicide or suicidal ideation in EHRs," in AMIA Annual Symposium Proceedings, vol. 2012, p. 1244, 2012.

[21] C. Hogan, J. Lunney, J. Gabel, and J. Lynn, "Medicare beneficiaries’ costs of care in the last year of life," Health Affairs, vol. 20, no. 4, pp. 188-195, 2001.

[22] J. Holland et al., "Service robots in the healthcare sector," Robotics, vol. 10, no. 1, p. 47, 2021.

[23] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," arXiv preprint arXiv:1904.05342, 2019.

[24] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," in Informatics for Health: Connected Citizen-Led Wellness and Population Health, pp. 246-250, 2017.

[25] N. Kang et al., "MedNLP: a multi-modal query-based system for cross-modal retrieval of medical cases," Journal of the American Medical Informatics Association, vol. 25, no. 5, pp. 512-519, 2018..

[26] J. Mahatpure, M. Motwani, and P. K. Shukla, "An electronic prescription system powered by speech recognition, natural language processing and blockchain technology," International Journal of Science & Technology Research (IJSTR), vol. 8, no. 08, pp. 1454-1462, 2019

[27] S. Maiti, "Extracting medical information from clinical text with NLP," Analytics Vidhya, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/02/extracting-medical-information-from-clinical-text-with-nlp/. [Accessed: 24-Sep-2023].

[28] A. Malte and P. Ratadiya, "Evolution of transfer learning in natural language processing," arXiv preprint arXiv:1910.07370, 2019.

[29] B. Manchanda and V. AnantAthavale, "Various Statistical Techniques Used in NLP," International Journal of Computer Applications & Information Technology, vol. 9, no. 1, p. 172, 2016.

[30] S. Manik, G. Singh, and R. Singh, "Design and analysis of stochastic DSS query optimizers in a distributed database system," Egyptian Informatics Journal, Cairo University, 2015.

[31] T. A. Manolio et al., "Implementing genomic medicine in the clinic: the future is here," Genetics in Medicine, vol. 15, no. 4, pp. 258-267, 2013.

[32] M. Marcus, "New trends in natural language processing: statistical natural language processing," Proceedings of the National Academy of Sciences, vol.

92, no. 22, pp. 10052-10059, 1995.

[33] A. K. McCallum, "MALLET: a machine learning for language toolkit," 2018. [Online]. Available: http://mallet.cs.umass.edu/. [Accessed: 24-Sep-2023].

[34] J. Meehan et al., "Precision medicine and the role of biomarkers of radiotherapy response in breast cancer," Frontiers in Oncology, vol. 10, p. 628, 2020.

[35] N. Menachemi and T. H. Collum, "Benefits and drawbacks of electronic health record systems," Risk Management and Healthcare Policy, pp. 47-55, 2011.

[36] L. Moja et al., "Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes: a randomized clinical trial," JAMA Network Open, vol. 2, no. 12, pp. e1917094-e1917094, 2019.

[37] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: fast and robust models for biomedical natural language processing," arXiv preprint arXiv:1902.07669, 2019.

[38] NLTK Project, "Natural Language Toolkit (NLTK)," 2018. [Online]. Available: https://www.nltk.org/. [Accessed: 22-Sep-2023].

[39] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli, "Exploring chemical space using natural language processing methodologies for drug discovery," Drug Discovery Today, vol. 25, no. 4, pp. 689-705, 2020.

[40] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in Proceedings of the Second Workshop on Shortcomings in Vision and Language, pp. 26-36, Jun. 2019.

[41] T. H. Payne, W. D. Alonso, J. A. Markiel, K. Lybarger, and A. A. White, "Using voice to create hospital progress notes: description of a mobile application and supporting system integrated with a commercial electronic health record," Journal of Biomedical Informatics, vol. 77, pp. 91-96, 2018..

[42] J. Pustejovsky and A. Stubbs, Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. O'Reilly Media, Inc., 2012.

[43] G. Randhawa, M. Ferreyra, R. Ahmed, O. Ezzat, and K. Pottie, "Using machine translation in clinical practice," Canadian Family Physician, vol. 59, no. 4, pp. 382-383, 2013.

[44] M. Rayner et al., "A methodology for comparing grammar-based and robust approaches to speech understanding," in INTERSPEECH, pp. 1877-1880, Sep. 2005.

[45] D. Riaño, M. Peleg, and A. Ten Teije, "Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges," Artificial Intelligence in Medicine, vol. 100, p. 101713, 2019.

[46] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," in Proceedings of the NIPS Workshop on Cost-Sensitive Learning, vol. 1, Dec. 2008.

[47] M. Sharma, G. Singh, R. Singh, and G. Singh, "Analysis of DSS queries using entropy based restricted genetic algorithm," Applied Mathematics & Information Sciences, vol. 9, no. 5, p. 2599, 2015.

[48] M. Sharma, G. Singh, R. Singh, and S. Singh, "Statistical Analysis of DSS Query Optimizer for a Five Join DSS Query," International Journal of Computer Applications, vol. 141, no. 6, pp. 1-4, 2016.

[49] M. Sharma, G. Singh, R. S. Virk, and G. Singh, "Design and comparative analysis of DSS queries in distributed environment," in 2013 International Computer Science and Engineering Conference (ICSEC), pp. 73-78, Sep. 2013.

[50] J. S. Son et al., "Association of blood pressure classification in Korean young adults according to the 2017 American College of Cardiology/American Heart Association guidelines with subsequent cardiovascular disease events," JAMA, vol. 320, no. 17, pp. 1783-1792, 2018.

[51] The Apache Software Foundation, "Stanford CoreNLP: natural language software," 2018. [Online]. Available: https://stanfordnlp.github.io/CoreNLP/. [Accessed: 22-Sep-2023].

[52] The University of Sheffield, "General Architecture for Text Engineering (GATE)," 2018. [Online]. Available: http://gate.ac.uk/. [Accessed: 24-Sep-2023].

[53] N. Vaci, D. Cocić, B. Gula, and M. Bilalić, "Large data and Bayesian modeling—aging curves of NBA players," Behavior Research Methods, vol. 51, pp. 1544-1564, 2019.

[54] R. Valencia-Garcia, R. Martinez-Bejar, and A. Gasparetto, "An intelligent framework for simulating robot-assisted surgical operations," Expert Systems with Applications, vol. 28, no. 3, pp. 425-433, 2005.

[55] X. Wang, G. Hripcsak, M. Markatou, and C. Friedman, "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study," Journal of the American Medical Informatics Association, vol. 16, no. 3, pp. 328-337, 2009.

[56] H. Wu et al., "SemEHR: A general-purpose semantic

search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research," Journal of the American Medical Informatics Association, vol. 25, no. 5, pp. 530-537, 2018.

[57] G. Xu, W. Rong, Y. Wang, Y. Ouyang, and Z. Xiong, "External features enriched model for biomedical question answering," BMC Bioinformatics, vol. 22, no. 1, p. 272, 2021.

[58] Ouerhani, N., Maalel, A., & Ben Ghézela, H. (2020). SPeCECA: a smart pervasive chatbot for emergency case assistance based on cloud computing. Cluster Computing, 23, 2471-2482.

Author Biography

Shashank Agarwal: Shashank Agarwal is a healthcare data science expert whose experience cuts across various areas in market access, artificial intelligence, brand analytics, predictive modeling, launch strategy, and multi-channel marketing in several Fortune 500 companies such as CVS Health, AbbVie, and IQVIA. Additionally, he holds a Master of Science in Engineering Management from The Johns Hopkins University, USA.

Praveen Gujar: Praveen Gujar is a seasoned Product Leader, excels in Digital Advertising, AI, and Cloud tech. With notable roles at LinkedIn, Twitter, and Amazon. He currently drives innovation and business growth as Director of Product at LinkedIn

Sriram Panyam: Sriram Panyam is an experienced engineering leader with a track record for developing technical organizations within major tech firms like Google, LinkedIn, and Amazon. His expertise spans large-scale systems, cloud platforms, AI/ML and data analytics.