Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 05 Issue: 06 | June-2025

From Structured to Unstructured: A Review on NLP Applications Transforming Healthcare Data

Aditya Narayan

[1]*Bangalore University*

---***---

Abstract - Healthcare data is produced at an unprecedented scale worldwide, consisting of both structured and unstructured components. Structured data such as lab values, vital signs, and diagnostic codes are traditionally easier to analyze, but they represent only a fraction of the total healthcare information. A vast majority of clinical data exists in unstructured formats, including free-text clinical notes, imaging reports, pathology narratives, and patient communications, which contain rich contextual and nuanced information. Natural Language Processing (NLP) has emerged as a critical technology for unlocking this underutilized resource by converting unstructured text into structured, computable data. This paper reviews the state-of-the-art NLP methodologies applied in healthcare, highlighting their role in transforming unstructured data into actionable knowledge. It explores key NLP techniques such as named entity recognition, relation extraction, sentiment analysis, and deep learning frameworks. The paper discusses diverse applications ranging from clinical documentation improvement and decision support to research and population health management. Challenges related to linguistic variability, domain adaptation, data privacy, and system integration are analyzed. Future prospects include real-time NLP, multilingual capabilities, and explainability, which promise to accelerate the integration of NLP-driven insights into clinical practice and research, ultimately enhancing patient care quality and operational efficiency.

Keywords- Natural Language Processing, Healthcare Data, Unstructured Data, Electronic Health Records, Clinical Text Analysis

Introduction

The healthcare industry is characterized by massive data generation from diverse sources such as hospitals, clinics, laboratories, imaging centers, and wearable devices. Electronic Health Records (EHRs) have digitized patient information, enabling data-driven healthcare delivery and research. However, despite advances in EHR adoption, a significant portion of valuable clinical data remains locked in unstructured text. Clinical narratives, including progress notes, discharge summaries, radiology interpretations, pathology reports, and patient-reported outcomes, capture complex information such as clinical reasoning, symptom descriptions, social determinants, and physician impressions, which are often not reflected in structured fields.

The complexity, variability, and ambiguity inherent in unstructured text make it difficult for traditional data analytics tools to extract meaningful insights. This gap hampers comprehensive patient profiling, accurate clinical decision-making, and population-level health analysis. Natural Language Processing (NLP), an AI subfield dedicated to enabling machines to understand and interpret human language, provides the technological framework to bridge this gap. By transforming free-text data into structured formats, NLP enables healthcare providers and researchers to unlock insights previously hidden in textual data, leading to improved clinical outcomes, optimized workflows, and accelerated research.

This paper offers an in-depth review of how NLP applications are revolutionizing the processing of healthcare data, emphasizing the transition from structured to unstructured data utilization. It covers foundational NLP techniques tailored to the medical domain, key applications transforming healthcare, integration strategies with existing clinical systems, and the challenges and future opportunities of this rapidly evolving field.

Fundamental Concepts and Techniques in NLP for Healthcare
Natural Language Processing involves multiple stages designed to process, analyze, and interpret text data effectively. At the foundational level, text preprocessing prepares raw clinical narratives by removing noise and standardizing content through tokenization (splitting text into words or phrases), lemmatization and stemming (reducing words to their root forms), and stop-word removal (eliminating commonly used words with little semantic value).

Named Entity Recognition (NER) is a core NLP task that identifies and classifies entities within clinical text, such as diseases, symptoms, medications, procedures, and anatomical locations. For example, an NLP system may identify "hypertension" as a disease entity or "aspirin" as a medication. However, clinical text is rife with complex terminology, abbreviations, and synonyms, which require sophisticated lexicons and domain-specific ontologies like the Unified Medical Language System (UMLS) to ensure accurate entity recognition.

Relation extraction follows entity recognition by identifying relationships between entities, such as drug-dosage, symptom-onset time, or disease-complication connections. This step is crucial for understanding clinical context and producing structured knowledge graphs from narrative text.

Negation detection algorithms play an essential role in healthcare NLP because clinical notes frequently mention the absence of symptoms or diseases, e.g., “no evidence of pneumonia.” Properly recognizing such negations prevents erroneous data extraction and misinterpretation.

Beyond these, advanced NLP models employ syntactic parsing to analyze sentence structure and semantic role labeling to determine the function of words in context. Machine learning approaches, particularly supervised learning using annotated corpora, have been widely adopted to train NLP systems to perform these tasks. In recent years, deep learning architectures such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based models (e.g., BERT and its biomedical adaptations like BioBERT) have significantly improved the ability of NLP systems to capture complex patterns and context in clinical text.

Applications Transforming Healthcare Data

NLP's ability to process unstructured data has led to its adoption across numerous healthcare domains, fundamentally changing how clinical data is captured, analyzed, and utilized.

One prominent application is clinical documentation improvement. NLP extracts critical clinical concepts and links them with appropriate medical codes (e.g., ICD-10, CPT codes), improving the accuracy of medical billing and reducing administrative burden. Automating coding processes saves clinicians significant time while enhancing revenue cycle management for healthcare providers.

Clinical Decision Support Systems (CDSS) are also increasingly leveraging NLP to incorporate insights from free-text clinical notes. By integrating structured and unstructured data, CDSS can generate more accurate alerts and recommendations, improving patient safety and care quality. For example, NLP can identify mentions of adverse drug reactions or potential drug interactions hidden within narrative notes, enabling early intervention.

In research and epidemiology, NLP accelerates the identification of patient cohorts for clinical trials or observational studies by extracting phenotypic data from unstructured clinical records. This capability expands the scale and scope of studies while reducing manual chart review workload. Additionally, NLP contributes to pharmacovigilance by mining adverse event reports and literature to detect drug safety signals.

Patient engagement has also benefited from NLP advancements. Chatbots and virtual health assistants employ NLP to interpret patient queries, provide symptom assessment, health education, medication reminders, and facilitate appointment scheduling, enhancing accessibility and adherence to treatment plans.

Integration with Electronic Health Records and Systems

The integration of NLP applications within Electronic Health Records (EHRs) is critical for operationalizing unstructured data insights in clinical workflows. Modern EHR platforms are increasingly embedding NLP modules to perform real-time extraction and summarization of clinical notes. This integration enables automated generation of problem lists, structured summaries, and documentation templates, reducing clinician workload and enhancing record completeness.

Interoperability is another vital aspect, where NLP tools are designed to work seamlessly across different EHR systems and data formats, fostering data sharing and aggregation for collaborative care and multicenter research. Standardized data models and APIs facilitate the incorporation of NLP-extracted information into clinical decision-making interfaces and reporting dashboards.

Challenges and Future Directions

Despite considerable progress, several challenges impede the widespread adoption of NLP in healthcare. The inherent variability of clinical language, including the use of abbreviations, misspellings, and evolving terminologies, poses significant difficulties for NLP algorithms. Annotated datasets essential for training robust models are scarce and expensive to create due to the need for domain expertise and privacy concerns.

Data privacy and security remain paramount, especially given the sensitive nature of health information. NLP applications must comply with regulatory frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe. Techniques like data de-identification and federated learning are being explored to mitigate privacy risks.

Furthermore, integrating NLP into existing clinical workflows demands usability considerations and clinician trust. NLP outputs must be accurate, explainable, and seamlessly presented to avoid alert fatigue or mistrust. Developing explainable AI methods that provide transparency into NLP decision processes is an active research area.

Future developments in healthcare NLP include the creation of multilingual models capable of processing diverse patient populations and healthcare settings. Advances in real-time NLP will allow immediate data extraction and clinical alerts at the point of care. The fusion of NLP with other modalities such as medical imaging and genomic data promises holistic patient insights.

Conclusion

Natural Language Processing is fundamentally transforming the landscape of healthcare data by enabling the effective use of unstructured clinical information. Through sophisticated algorithms and integration into healthcare systems, NLP converts rich narrative texts into structured, actionable data that supports clinical decision-making, enhances documentation, accelerates research, and improves patient engagement. Although challenges remain in terms of language complexity, data privacy, and workflow integration, ongoing technological advancements and collaborative efforts are driving the maturation of NLP solutions. As these tools continue to evolve and integrate more deeply into clinical practice, NLP will become indispensable in achieving a data-driven, patient-centered healthcare ecosystem, unlocking new potentials in personalized medicine and population health management.

References

- Nancy, A. M., & Maheswari, R. (2020). A review on unstructured data in medical data. *J. Crit. Rev*, *7*, 2202-2208.
- Chinthala, L. K. (2021). Future of supply chains: Trends in automation, globalization, and sustainability. *International Journal of Scientific Research & Engineering Trends*, *7*(6), 1-10.
- Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science*, *8*(8), 44-50.
- Kolla, V. R. K. (2021). Cyber security operations centre ML framework for the needs of the users. International Journal of Machine Learning for Sustainable Development, 3(3), 11-20.
- Kolla, V. R. K. (2020). India's Experience with ICT in the Health Sector. Transactions on Latest Trends in Health Sector, 12, 12.
- Kolla, V. R. K. (2016). Forecasting Laptop Prices: A Comparative Study of Machine Learning Algorithms for Predictive Modeling. International Journal of Information Technology & Management Information System.
- Kolla, V. R. K. (2021). Prediction in Stock Market using AI. Transactions on Latest Trends in Health Sector, 13, 13.
- Kolla, Venkata Ravi Kiran, Analyzing the Pulse of Twitter: Sentiment Analysis using Natural Language Processing Techniques (August 1, 2016). International Journal of Creative Research Thoughts, 2016, Available at SSRN: https://ssrn.com/abstract=4413716
- Chinthala, L. K. (2021). Diversity and inclusion: The business case for building more equitable organizations. *Journal of Management and Science*, *11*(4), 85-87. Retrieved from https://jmseleyon.com/index.php/jms/article/view/834
- Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., ... & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, *73*, 14-29.
- Chinthala, L. K. (2018). Environmental biotechnology: Microbial approaches for pollution remediation and resource recovery. In Ecocraft: Microbial Innovations (Vol. 1, pp. 49–58). SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5232415
- Adnan, K., Akbar, R., Khor, S. W., & Ali, A. B. A. (2020). Role and challenges of unstructured big data in healthcare. *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 1*, 301-323.
- Chinthala, L. K. (2018). Fundamentals basis of environmental microbial ecology for biofunctioning. In Life at ecosystem and their functioning. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5231971
- Mugisha, C., & Paik, I. (2020, December). Pneumonia outcome prediction using structured and unstructured data from EHR. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2640-2646). IEEE.
- Chinthala, L. K. (2017). Functional roles of microorganisms in different environmental

processes. In Diversified Microbes (pp. 89–98). SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5232387

- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering*, *2018*(1), 4302425.
- Chinthala, L. K. (2016). Environmental microbiomes: Exploring the depths of microbial diversity. In Microbial Ecology: Shaping the Environment (Vol. 2, pp. 1–12). SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5232403
- Wulff, A., Mast, M., Hassler, M., Montag, S., Marschollek, M., & Jack, T. (2020). Designing an openEHR-based pipeline for extracting and standardizing unstructured clinical data using natural language processing. *Methods of information in medicine*, *59*(S 02), e64-e78.
- Chinthala, L. K. (2015). Microbes in action: Ecological patterns across environmental gradients. In Impact of microbes on nature (pp. 45–56). SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5232016
- Kumar, A. NLP Techniques and Applications in Healthcare Systems. In *Ambient Assisted Living (AAL) Technologies* (pp. 225-239). CRC Press.
- Chinthala, L. K. (2014). Dynamics and applications of environmental microbiomes for betterment of ecosystem. In Role of microbiomes in society PhDians (pp. 1–13). SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5231959
- Adnan, K., Akbar, R., & Khor, S. W. (2019). Role and Challenges of Unstructured. *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 1*, *1042*, 301.
- Chinthala, L. K. (2021). Business in the Metaverse: Exploring the future of virtual reality and digital interaction. *International Journal of Science, Engineering and Technology*, *9*(6). ISSN (Online): 2348-4098, ISSN (Print): 2395-4752.
- Shah, R. F., Bini, S., & Vail, T. (2020). Data for registry and quality review can be retrospectively collected using natural language processing from unstructured charts of arthroplasty patients. *The Bone & Joint Journal*, *102*(7 Supple B), 99-104.
- Chinthala, L. K. (2021). Revolutionizing business operations with nanotechnology: A strategic perspective. *Nanoscale Reports*, *4*(3), 23-27.
- Hong, N., Wen, A., Shen, F., Sohn, S., Liu, S., Liu, H., & Jiang, G. (2018). Integrating structured and unstructured EHR data using an FHIR-based type system: a case study with medication data. *AMIA Summits on Translational Science Proceedings*, *2018*, 74.