Streamlining Healthcare with NLP and AI: Extracting Medical Information from Unstructured Text and Linking of Medical Codes

Raghavendra Ganiga[1,2], Youngwoo Oh[1] and Wooyeol Choi[1,*]

[1]Chosun University, [2]Manipal Academy of Higher Education (MAHE)

raghavendra_ganiga@chosun.ac.kr, ywo@chosun.kr, wyc@chosun.ac.kr

Abstract

Electronic health records (EHRs) are a rich source of real-world data that can be utilized by healthcare providers and researchers for clinical decision-making. EHRs are only partially utilized due to the challenges associated with automated data extraction. Health information is available in both structured and unstructured formats. However, clinicians find it challenging to extract meaningful insights from unstructured data due to its complex and time-consuming nature. Natural language processing (NLP), a form of artificial intelligence (AI), has the potential to revolutionize patient care by providing more efficient and accurate data analysis. This research proposes an NLP machine-learning model that can generate clinically relevant synthetic data to assist clinicians in their practice. The results suggest that the NLP system is a promising approach for extracting valuable insights from the vast amounts of unstructured data in EHR.

I. Introduction

Electronic health records (EHRs) provide a wealth of information about an individual's health, including demographics, diagnoses, medications, laboratory results, and waveform signals [1]. The digitization of healthcare through information and communication technologies (ICTs) has facilitated the development of EHRs, which have the potential to improve diagnosis and treatment outcomes for various diseases. EHRs, along with other patient documentation systems, offer a rich resource of data for advancing our understanding of disease conditions, complementing traditional study designs. The integration of ICTs in healthcare can lead to more comprehensive views of patients' health status and more efficient and effective treatments, ultimately improving patient outcomes [2].

Many hospitals are adopting AI technologies to improve the efficiency of their operations and patient management [3, 4]. EHRs data can be used in various healthcare sectors, such as disease management, new drug development, treatment planning, etc. One of the key benefits of using AI is the ability to predict the occurrence of specific diseases or provide personalized treatment by classifying individualized patient characteristics [5]. In developed countries, there is a growing interest among researchers in structuring and standardizing medical data to use it clinically. In addition, machine learning algorithms can be applied to EHR data to create predictive models that can help healthcare providers identify patients at risk of specific diseases or predict treatment outcomes. Standardizing medical data using coding systems, such as ICD-10, SNOMED-CT, and LOINC, is essential for the effective use of EHR data [6]. By using NLP techniques to extract information and apply standardized codes to EHR data, healthcare providers and researchers can more easily access and analyze large volumes of patient data [7][8].

II. Proposed Method

NLP is a critical tool for extracting meaningful insights from clinical documents. Despite the challenges presented by the under-utilized free text data in narrative clinical documents, transforming this data into actionable knowledge requires systematic approaches. The scispaCy package for Python provides reliable and usable models for processing biomedical and scientific text [9]. The most impressive aspects of spaCy are its neural network models for tagging, parsing, named entity recognition (NER), text classification, and various other applications.

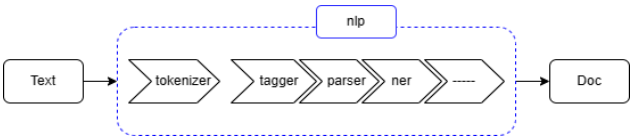

Figure 1. Language processing pipeline.

The language processing pipeline depicted in Figure 1 is crucial for extracting meaningful insights from unstructured clinical documents. It consists of several components executed in the correct order, including a tokenizer, tagger, parser, named entity recognizer (NER), and more. To ensure accurate and reliable extracted data, preprocessing of unstructured data is essential. This involves removing noise, standardizing the format, and normalizing terminology. After preprocessing, NLP techniques such as named entity recognition, information extraction, and sentiment analysis can be applied to extract relevant information and identify relationships between concepts. The pre-built components and user-created components of spaCy can be added to the pipeline using the Language.add_pipe method. The model was tested using a Kaggle dataset consisting of medical transcription

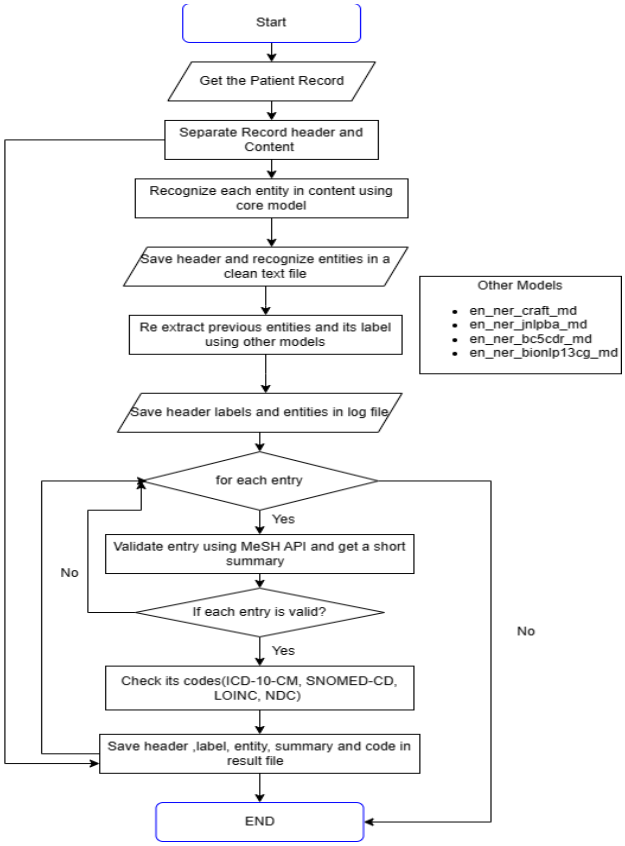

Figure 2. Proposed NLP machine learning model.

samples from various specialties [10]. A total of 40 patient records were selected for testing purposes.

The proposed methodology for using NLP to extract insights from EHRs involves several steps, as shown in Figure 2. To extract insights from EHRs using NLP, the unstructured clinical text data must be preprocessed. This involves obtaining the data from EHRs, separating header information, and using the Spacy core model to recognize each entity. The core model includes tokenization, part-of-speech tagging, and named entity recognition. The scispaCy package offers specialized models for processing clinical, scientific, and biomedical text data, which is useful for healthcare professionals and researchers.

Entities and their labels are re-extracted using models with header labels and entities saved in a log file. The medical subject headings (MeSH) API is then used to validate the entry and obtain a short summary of the patient details. The Mesh API is used to validate each entity and obtain a short summary of the patient details. MeSH is a controlled vocabulary used in biomedical research to classify and organize articles and books. The MeSH API validates each entity and provides a summary of patient details. Diseases, medications, and lab tests are mapped to standardized codes like ICD-10, SNOMED-CT, and LOINC. The results are saved in CSV format, enabling easy loading into a database or web application. The structured data provides meaningful insights for healthcare professionals and researchers to enhance patient care and outcomes. The overall performance of the NLP system can be evaluated by measuring its accuracy, precision, recall, F1 score, and other performance metrics on a test dataset or by conducting user studies with medical experts. These analyses can help improve the accuracy and effectiveness of the NLP system and make it more useful for researchers, doctors, and patients in managing and understanding diseases.

The study findings indicate that the NLP system is a promising approach for extracting valuable insights from the vast amount of unstructured data present in EHR notes. Nevertheless, further research is required to enhance the model's performance by integrating new approaches to extract insights from clinical notes.

III. Conclusion

NLP offers significant opportunities to improve evidence-based decision-making in public health by extracting valuable insights from unstructured data. The proposed NLP model takes unstructured medical text as input from various sources, extracts medical conditions, and links them to specific standardized medical codes. This promising approach can inform clinical decision-making, remote monitoring, self-care, and medication adherence for chronic disease management.

ACKNOWLEDGMENT

This work was supported by the Technology development Program(S3312532) funded by the Ministry of SMEs and Startups(MSS, Korea).

REFERENCES

[1] J. G. Richter and C. Thielscher, “New developments in electronic health record analysis,” Nature Reviews Rheumatology, vol. 19, no. 2, pp. 74–75, 2023.

[2] C. Y. Hui, A. Abdulla, Z. Ahmed, H. Goel, G. Monsur Habib, M. Nurmansyah et al., “Mapping national ICT infrastructure to the requirements of potential digital health interventions in low-and middle-income countries.” Journal of global health, vol. 12, 2022.

[3] M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen, “Challenges and opportunities beyond structured data in analysis of electronic health records,” Wiley Interdisciplinary Reviews: Computational Statistics, vol. 13, no. 6, p. e1549, 2021.

[4] D. Kotecha, F. W. Asselbergs, B. Casadei et al., “Code-ehr best practice framework for the use of structured electronic healthcare records in clinical research,” European heart journal, vol. 43, no. 37, pp.3578–3588, 2022.

[5] J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui, “A survey of deep learning for electronic health records,” Applied Sciences, vol. 12, no. 22, p. 11709, 2022.

[6] M. M. Pai, R. Ganiga, R. M. Pai, and R. K. Sinha, “Standard electronic health record (ehr) framework for indian healthcare system,” Health Services and Outcomes Research Methodology, vol. 21, no. 3, pp. 339–362, 2021.

[7] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nature medicine, vol. 25, no. 1, pp. 44–56, 2019.

[8] F. Liu, C. Weng, and H. Yu, “Advancing clinical research through natural language processing on electronic health records: traditional machine learning meets deep learning,” Clinical Research Informatics, pp. 357–378, 2019.

[9] G. B. Negrini, Biomedical text natural language processing (bionlp) using scispacy, (https://spacy.io/usage).

[10] T. Boyle, Medical transcriptions, (https://www.kaggle.com/datasets/).