DOI: 10.53469/jcmp.2024.06(08).67

# Structured Data Extraction of Unstructured Clinical Notes Based on Natural Language Processing

**Arjit Amol More**

Sipna College of Engineering and Technology, Amravati, Maharashtra, India
*arjit@sipnaengg.ac.in*

**Abstract:** ***Electronic Health Records (EHRs) are pivotal in modern healthcare, housing a treasure trove of patient information. They are real-time, patient-centered records that make information available instantly and securely to authorized users. However, a substantial portion of this data resides in unstructured clinical notes, presenting significant challenges for data extraction and utilization. This research paper investigates the issues posed by unstructured clinical notes application of Natural Language Processing (NLP) techniques in the healthcare sector to extract structured patient data from unstructured clinical notes. By utilizing NLP algorithms, healthcare institutions can unlock invaluable insights within EHRs, leading to improved patient care, clinical research, and administrative efficiency. This paper addresses various NLP approaches, the implementation of pre-trained SpaCy and Med7Modelfor extracting structural data, and the potential for future advancements in this critical area of healthcare informatics.***

**Keywords:** Electronic health records, Unstructured Clinical Notes, Natural Language Processing, patient data extraction, SpaCy, healthcare informatics.

## 1. Introduction

An electronic version of a patient's medical history is called an electronic health record, or EHR. Electronic Health Records (EHRs) are databases created during clinical events or interactions. Demographics, issues, prescriptions, progress notes, and other important administrative clinical data can be included. One major distinction between an EMR and an EHR is the instantaneous sharing of all information. An EMR captures information from a single healthcare provider and is only accessible by that single healthcare provider. Nonetheless, EHRs are made to be utilized by a variety of healthcare organizations and providers [1]-[4]. EHRs can be shared throughout various healthcare settings and are updated over time by providers. These patient-centered, real-time records give authorized users instantaneous, secure access to information. Within EHRs, clinical notes offer a narrative account of patient conditions, treatments, and progress. A necessary format for recording an interaction with a patient is a clinical note. They play a crucial role in the medical records of patients and can significantly affect the kind of care they receive. However, most of these notes are unstructured free-text documents. Unstructured clinical notes within EHRs comprise diverse data types, including progress notes, discharge summaries, consultation reports, and medical histories. They are typically composed in natural language, which presents a challenge for data extraction due to the variability in language and writing styles [7].

Unstructured clinical notes introduce several challenges, including data noise, linguistic diversity, data redundancy, and the need to ensure patient data privacy and security. Data noise may include irrelevant or duplicated information, making it essential to identify and filter the pertinent data accurately. This Paper aims to explore the application of Natural Language Processing (NLP) techniques to extract structured patient data from unstructured clinical notes within EHRs. The primary objectives are as follows:

1) Investigate the challenges posed by unstructured clinical notes.
2) Review the critical role of NLP in healthcare informatics.
3) Provide implementation and practical examples of NLP applications for patient data extraction.

## 2. Unstructured issues in Clinical notes

- A patient's medical history, symptoms, diagnosis, course of treatment, and progress are all described in detail in their clinical notes, which are an integral element of the medical record. It offers a chronology of the patient's medical occurrences and interventions and includes both subjective and objective data organized chronologically [3]. Unstructured issues in clinical notes within Electronic Health Records refer to challenges associated with free-Text narrative entries that lack a standardized format. Following are some common issues:
- **Free-Text Variability**: Clinicians may use diverse language, abbreviations, or acronyms, making it difficult to extract and interpret information consistently.
- **Lack of Standardization**: There may be inconsistencies in how different healthcare professionals document information, leading to variations in terminology, style, and content.
- **Limited Searchability**: Unstructured data can be challenging to search and retrieve compared to structured data. This can impede quick access to relevant information.
- **Semantic Ambiguity**: The meaning of terms or phrases may vary between users or over time, leading to potential misinterpretation and miscommunication.
- **Contextual Understanding**: Understanding the context of certain statements or notes may be difficult, especially when details are scattered across multiple unstructured entries.
- **Difficulty in Data Mining**: Analyzing unstructured data for research or quality improvement purposes is more complex compared to structured data.

- **Incomplete Information**: Clinicians may not provide exhaustive details in free-text notes, leaving gaps in the patient's medical history or current condition.
- **Inability to Capture Structured Data**: Information that could be better represented in a structured format may end up in free-text notes, reducing the potential for interoperability and data exchange.
- **Redundancy and Repetition**: Clinicians may repetitively document certain information in different parts of the record, leading to redundancy and potential confusion.
- **Limited Decision Support**: Extracting actionable insights or implementing decision support systems can be challenging when dealing with unstructured data compared to structured data.
- **Legal and Ethical Concerns**: Ensuring compliance with privacy and security regulations becomes more challenging when dealing with unstructured data, especially if sensitive information is not clearly defined. [2,3]
- In order to tackle these unstructured problems, advanced natural language processing (NLP) algorithms are being developed and put into use, as well as standardized terminology being promoted described in the next Section.

## 3. Extraction of structured data through Natural Language Processing

### 3.1 Title and authors

NLP is a multidisciplinary field focusing on computer and human language interaction. In healthcare, NLP plays a vital role in data processing, enabling the extraction of structured data from unstructured text.NLP holds the promise of addressing the challenges posed by unstructured clinical notes. Following figure shows the steps performed for extracting structured data from unstructured clinical Notes[6]-[9].

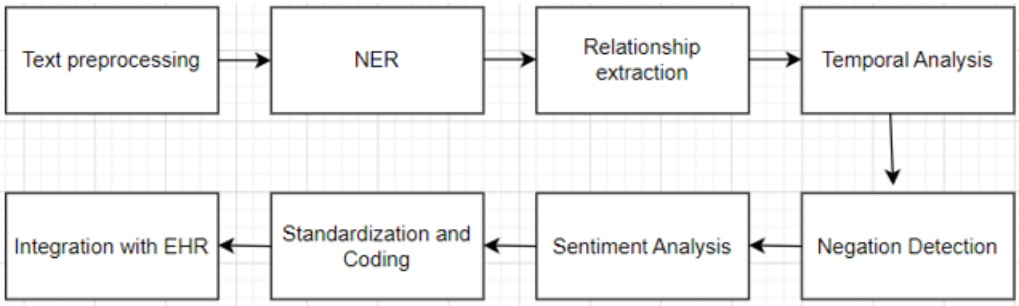


**Figure 1:** Steps followed for Extraction of structured data

**Step-1: Text Preprocessing**:
NLP algorithms start by preprocessing the unstructured text. This involves tasks such as tokenization (breaking text into words or phrases), stemming (reducing words to their root form), and removing stop words (common words that do not carry much meaning). This step helps in preparing the text for more advanced analysis.

**Step-2: Named Entity Recognition (NER):**
NER is a crucial task in extracting structured data from clinical notes. It involves identifying and classifying entities mentioned in the text, such as patient names, medical conditions, medications, procedures, and dates. NLP models trained for healthcare can recognize and categorize these entities accurately.

**Step-3: Relationship Extraction:**
After identifying entities, NLP systems work on understanding the relationships between them. For example, linking a medication to a specific medical condition or associating a procedure with a date. This step is essential for building a coherent and structured representation of the information contained in clinical notes.

**Step-4: Temporal Analysis:**
Clinical notes often include temporal information, such as the onset of symptoms, dates of procedures, or changes in medication. NLP models can analyze the temporal aspect of the text to order events chronologically, providing a timeline of the patient's medical history.

**Step-5: Negation Detection:**
Negation detection is crucial for understanding the context of information in clinical notes. NLP models can identify instances where a condition or treatment is explicitly negated, helping to avoid misinterpretation of information.

**Step-6: Sentiment Analysis:**
Sentiment analysis may be applied in certain cases to understand the tone or context of the clinical notes. This can be especially important in mental health settings or when assessing the impact of a treatment on a patient's well-being.

**Step-7: Standardization and Coding:**
Once the relevant information is extracted and understood, NLP systems can standardize the data by mapping it to standardized medical codes such as SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms) or LOINC (Logical Observation Identifiers Names and Codes). This facilitates interoperability and integration with existing healthcare systems.

**Step-8: Integration with Electronic Health Records (EHRs):**
The final structured data, extracted and processed by NLP, can be integrated into electronic health records. This enhances the accessibility and usability of the information for healthcare professionals, supporting better clinical decision-making.

## 4. Initial Experiment performed using SpaCy model for data extraction

Creating a complete and accurate system for extracting structured data from unstructured clinical notes is a complex task that may require a more sophisticated approach, potentially involving specialized models and training on medical data[10][11]. Here is a more comprehensive example using spaCy and the Med7 model, a pre-trained spaCy model for extracting medication-related information. spaCy is an open-source library designed for natural language processing (NLP) in Python.written in Cython, making it fast and capable of processing large amounts of text quickly. It provides efficient and easy-to-use tools for various NLP tasks, such as tokenization, part-of-speech tagging, named entity recognition, and more.spaCy comes with pre-trained models for various languages, allowing users to perform common NLP tasks without having to train models from scratch.The Med7 model is a pre-trained spaCy model specifically designed for extracting medication-related information from clinical text.

```
import spacy
# Load spaCy Med7 model
nlp = spacy.load("en_core_med7")

def extract_medication_information(text):
 # Process the text using spaCy Med7 model
   doc = nlp(text)
 # Initialize variables to store extracted medication information
   medications = []
 # Extract medication entities
   for ent in doc.ents:
      if ent._.is_medication:
         medications.append({'medication': ent.text, 'start': ent.start_char, 'end': ent.end_char,
  'route': ent._.route})

   return medications

# Example usage
clinical_note = "The patient was prescribed 10mg of Lipitor to be taken orally once a day."

medications = extract_medication_information(clinical_note)

print("Medication Information:")
for medication in medications:
   print(medication)
```

## Result

The code processes the clinical note and extracts medication-related information using the spaCy Med7 model. The extracted medication information includes the name of the medication, the start and end positions in the text, and the route of administration

```
Medication Information:
{'medication': '10mg of Lipitor', 'start': 29, 'end': 46, 'route': 'orally once a day'}
# Example usage
Clinical_note= "Arjun was prescribed 375 mg of Clavam with vitazinc to be taken orally twice a day."

Medication Information:
{'medication': '375 mg of Clavam with vitazinc', 'start': 26, 'end': 63, 'route': 'orally twice a day'}
```

## 5. Conclusion

In conclusion, the application of Natural Language Processing for extracting structured patient data from unstructured clinical notes within EHRs is a transformative advancement in healthcare informatics. In this paper an accurate system using Spacy Model for extracting structured data from unstructured clinical notes is implemented and tested with some unctructured data discussed in section 4. NLP has demonstrated its ability to address the challenges associated with unstructured data, resulting in improved patient care, enhanced clinical research, and increased administrative efficiency. As NLP technology continues to evolve, its role in healthcare will become increasingly central, ensuring that valuable patient data within clinical notes is utilized to its full potential.

## References


[1] Beasley JW, Holden RJ, Sullivan F. Electronic health records: research into design and implementation. Br J Gen Pract. 2011 Oct;61(591):604-5. doi: 10.3399/bjgp11X601244. PMID: 22152827; PMCID: PMC3177114.

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8735614/

[3] Su, Yu-Hsiang and Chao, Ching-Ping and Hung, Ling-Chien and Sung, Sheng-Feng and Lee, Pei-Ju," A Natural Language Processing Approach to Automated Highlighting of New Information in Clinical Notes",Applied Sciences,2020, doi = 10.3390/app10082824

[4] https://towardsdatascience.com/clinical-natural-language-processing-5c7b3d17e137

[5] https://www.analyticsvidhya.com/blog/2023/02/extracting-medical-information-from-clinical-text-with-nlp/

[6] V. G, H. R and J. Hareesh, "Relation Extraction in Clinical Text using NLP Based Regular Expressions," **2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)**, Kannur, India, 2019, pp. 1278-1282, doi: 10.1109/ICICICT46008.2019.8993274.

[7] D. Wang, J. Su and H. Yu, "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language," in **IEEE Access**, vol. 8, pp. 46335-46345, 2020, doi: 10.1109/ACCESS.2020.2974101.

[8] A. Q. Mahlawi and S. Sasi, "Structured data extraction from emails," **2017 International Conference on Networks & Advances in Computational Technologies (NetACT)**, Thiruvananthapuram, India, 2017, pp. 323-328, doi: 10.1109/NETACT.2017.8076789.

[9] H. Déjean, "Extracting structured data from unstructured document with incomplete

resources," **2015 13th International Conference on Document Analysis and Recognition (ICDAR)**, Tunis, Tunisia, 2015, pp. 271-275, doi: 10.1109/ICDAR.2015.7333766.

[10] D. Kumar, S. Pandey, P. Patel, K. Choudhari, A. Hajare and S. Jante, "Generalized Named Entity Recognition Framework," **2021 Asian Conference on Innovation in Technology (ASIANCON)**, PUNE, India, 2021, pp. 1-4, doi: 10.1109/ASIANCON51346.2021.9544652.

[11] Varun Reji , Srikanth Reddy N , Umar Shariff,"INFORMATION EXTRACTION USING NATURAL LANGUAGE PROCESSING",International Journal of Scientific Research in Engineering and Management (IJSREM),Vol. 06,2022,doi:10.55041/IJSREM13271.