Healthcare Data Summarization via Medical Entity Recognition and Generative AI

Giuseppe Riccio[1,2], Antonio Romano[1,2], Andriy Korsun[1,2], Michele Cirillo[1,2], Marco Postiglione[1,2], Valerio La Gatta[1,2], Antonino Ferraro[1,2], Antonio Galli[1,2] and Vincenzo Moscato[1,2]

[1]*University of Naples "Federico II", Via Claudio 21, Naples, 80125, Italy*
[2]*BIG DATA CINI National LAB - Node University of Naples "Federico II", Naples, 80125, Italy*

Abstract

This paper presents a fully automated approach for extracting value from content that lies hidden in Electronic Health Records (EHRs) using Large Language Models (LLMs) and Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER) and Entity Linking (L). In particular, the state-of-the-art approaches used to solve this task suffer from problems related to poor automation, given the laborious process of fine-tuning the models used and the difficult interpretation of the results obtained from them. The solution proposed in this work, on the other hand, aims to show the potential of NLP and generative AI to extract the relevant medical concepts contained within EHRs and generate a summary of the entire clinical history of each patient to construct a simple and intuitive dashboard that supports medical personnel with relevant medical information and useful analytics in order to diagnose and make decisions regarding the clinical condition of a patient.

Keywords

Named Entity Recognition, Entity Linking, Relation Extraction, Summarization, Generative AI

1. Introduction

The exponential increase of complex aggregated data in the healthcare sector and beyond has made understanding such data by medical professionals a challenging task. Extracting relevant information from Electronic Health Records [1], such as clinical notes, represents a significant challenge that requires advanced solutions based on the potential of Big Data and Natural Language Processing (**NLP**). In this article, we review the current approaches proposed in the scientific literature and present a possible solution that takes advantage of Named Entity Recognition (**NER**), Entity Linking (**L**), Relation Extraction (**RE**) and text synthesis techniques such as Large Languages Models (**LLMs**).

ITADATA2023, https://www.itadata.it/2023/calls, June 22, 2023
giuseppe.riccio9@studenti.unina.it (G. Riccio); antonio.romano45@studenti.unina.it (A. Romano); a.korsun@studenti.unina.it (A. Korsun); michele.cirillo2@studenti.unina.it (M. Cirillo); marco.postiglione@unina.it (M. Postiglione); valerio.lagatta@unina.it (V. L. Gatta); antonino.ferraro@unina.it (A. Ferraro); antonio.galli@unina.it (A. Galli); vmoscato@unina.it (V. Moscato)
https://github.com/giuseppericcio (G. Riccio); https://github.com/LaErre9 (A. Romano); https://github.com/andriykorsun (A. Korsun)
0009-0002-8613-1126 (G. Riccio); 0009-0000-5377-5051 (A. Romano)

In the scientific literature, several approaches have been proposed for the automatic extraction and synthesis of information from clinical records. One of the main approaches used is Named Entity Recognition (NER) [2], that focuses on the identification and extraction of relevant entities, which in this case could include diseases, drugs, medical procedures and symptoms within clinical texts. The use of clustering techniques [3] is common to organise and categorise extracted data, e.g. by applying clustering algorithms to group together notes dealing with similar topics. Furthermore, text synthesis techniques are used through the use of Generative Language Models [4], which enable the generation of coherent and contextually appropriate summaries based on the extracted data, for example, reports on a patient's medical history, providing a clear and concise picture of medical conditions, prescribed drugs and relevant procedures performed.

This paper presents a summary of the final project conducted as part of the Big Data Engineering course offered in the Master's Degree program in Computer Science at the University of Naples Federico II. The project aimed to address the challenges of the Big Data domain, with a specific focus on data management, processing, analysis, report generation, and protection. Our proposal is based on the combination of advanced techniques previously discussed, such as NER and LLMs.

To evaluate the effectiveness of our solution, we employed the MIMIC III dataset as our data source. This dataset is widely recognized for its extensive coverage, representativeness, and richness of clinical information. Through the application of our approach to this dataset, we demonstrate the information extraction and synthesis process, presenting the obtained results and their validity in the clinical context.

With this article, we fill a significant gap in the scientific literature by making a relevant contribution to the development of advanced approaches for the automatic extraction and synthesis of concepts from medical records, exploiting the potential of Big Data and LLMs to provide a comprehensive view of a patient's medical condition over time.

The results obtained from this study are valuable both for the academic community and the industry, as they offer a solid foundation for further research and advancements in the field of clinical data management and data engineering.

2. Related Work

In other fields than biomedical, several annotation interfaces have been developed for popular Natural Language Processing (NLP) tasks, such as Named Entity Recognition (NER), Entity Linking (EL), Relation Extraction (RE), Entity Normalization, Dependency Parsing, Chunking and so on. Among the available options, open-source tools such as **BRAT** (Stenetorp et al., 2012 [5]) have gained popularity. BRAT not only facilitates the management, monitoring and collection of annotated document corpuses, but also supports general annotation tasks. Another tool, **Prodigy**[1], is a commercial product that offers a modern annotation method for creating training and evaluation data for machine learning models. Although this tool can use various models to suggest entities, being based on the well-known NLP **SpaCy** library, it lacks automated integration with existing biomedical NER+L systems. Regarding biomedical NER+L, previous

[1]Documentation available at this site: https://prodi.gy/docs

scientific research has introduced tools such as **MetaMAP** (Aronson, 2001 [6]) and **CTakes** (Savova et al., 2010 [7]). These tools allow users to inspect recognized entities but do not provide mechanisms to correct and refine concepts or specify additional annotations based on specific research areas. Another tool called **SemEHR** (Wu et al., 2018 [8]) focuses on biomedical NER+L but differs in its approach from previous tools. Indeed, SemEHR allows for the incorporation of customized preprocessing and postprocessing steps and supports research-specific use cases. However, it does not directly improve the NER+L model through an interface, but treats the provided NER+L model as a black-box model, with no possibility of changing the recognized entities or obtaining more in-depth meta-information on those returned.

Regarding the summarization task, there are two main approaches in the literature: the first, called extractive summarization, involves the generated summary being composed of sentences extracted from the text provided as input based on a metric of importance of those sentences in the context of the text. The second approach, called abstractive summarization, involves extracting words within the text and reprocessing them to compose semantically related sentences. Regarding extractive summarization, several solutions have been proposed involving the use of neural networks, in which the problem is formulated as a classification task and networks composed of encoders and decoders are used (Cheng and Lapata, 2016 [9]; Nallapati et al., 2016 [10]) or pre-trained language models (Egonmwan and Chali, 2019 [11]; Liu and Lapata, 2019 [12]). In contrast, the state of the art in abstractive summarization involves numerous approaches, the most popular of which is the use of pre-trained encoder-decoder transformer models on a masked pre-training input target, the most popular of which are **MASS** (Song et al., 2019 [13]), **UniLM** (Dong et al., 2019 [14]), and **BART** (Lewis et al., 2019 [15]).

However, the approaches just mentioned require manual work to classify each document and the concepts it contains, and it is impractical to manually annotate large datasets such as those of patient notes. In this paper we try to explore, therefore, not only the development of a simple interface for annotating clinical texts, through NER+L models but also the generation of complete summaries of a patient's clinical history using LLMs, which allows to provide these summaries starting from the clinical notes treated in an appropriate way in a rapid and completely automated way with no need to fine tune a model as required by other approaches. Furthermore, through the integration of the entities extracted from the NER+L+RE and with the patient's summary, it is possible to provide, through a simple and intuitive dashboard, a series of analytics that support the diagnoses and decisions to be made with respect to an ill patient by competent medical personnel.

3. Methodological Workflow

Within our study, therefore, our goal was to develop a solution for the automatic summarization and report generation of clinical notes using, as previously written, a combination of Named Entity Recognition (NER), Linking (L), Relation Extraction (RE) and Language Models (LLMs). In order to achieve this, we followed a detailed workflow as follows: (Figure 1)

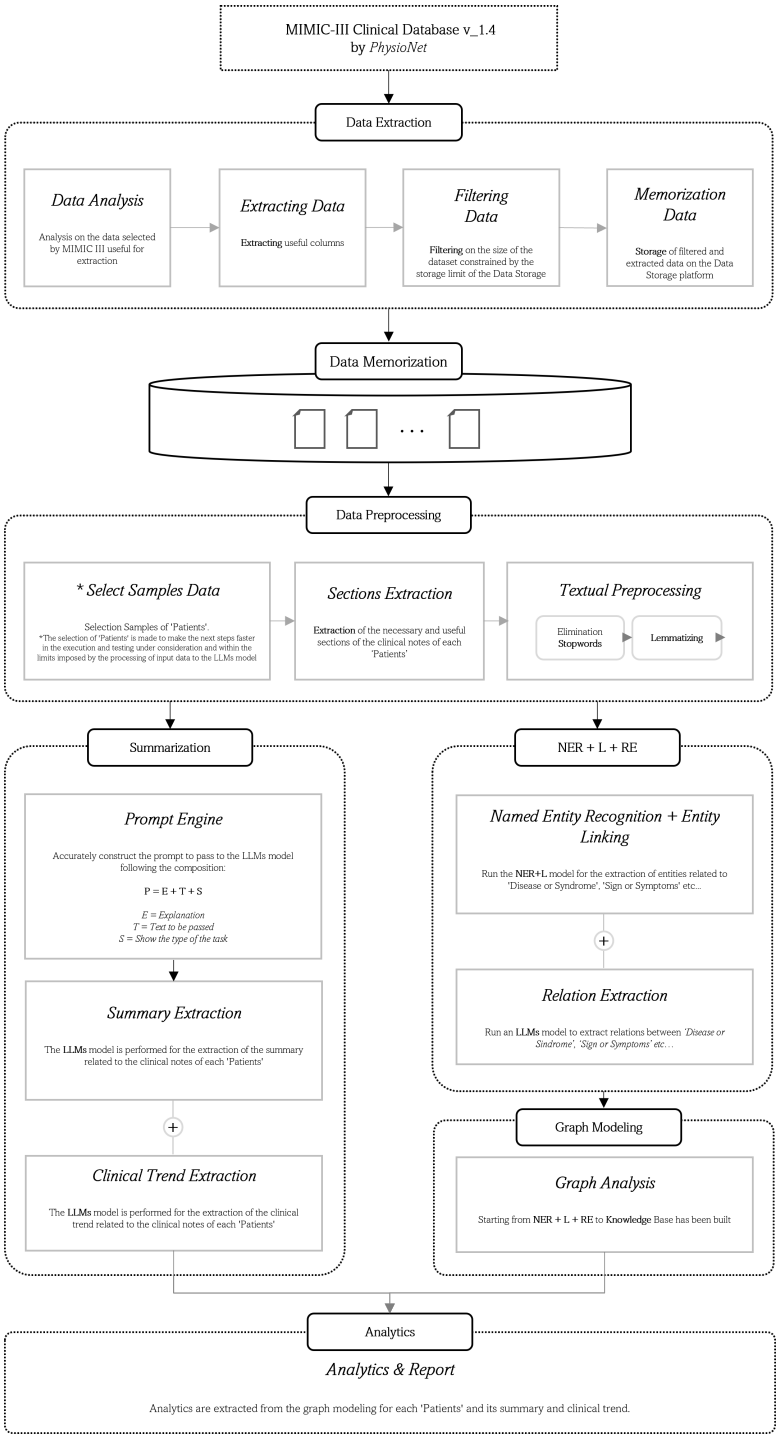


Figure 1: Workflow of the proposed method

3.1. Data collection

We used the anonymised clinical database provided by **MIMIC-III** [16] [17] [18]. In particular, we extracted information from the tables "PATIENTS" and "NOTEEVENTS". This data collection phase forms the basis of our study, although further processing is necessary to improve the data quality.

3.2. Data extraction

We conducted a data analysis to understand the distribution and composition of the clinical notes. This analysis helped us decide which fields to extract. During the data extraction process, we applied additional filters according to the limitations imposed by the selected data storage. In particular, we set a maximum limit of 10,000 tokens for the sum of the documents. We performed additional filtering to select at least 2 documents per patient and removed duplicate rows. Moreover, we extracted the demographic information of the patients from the "PATIENTS" table and merged it with the clinical notes dataframe to obtain a final dataframe upon which to perform subsequent project operations. The final data was stored in the selected data storage.

3.3. Data preprocessing

We selected a sample of the previously extracted data in order to make the further processes more efficient. Next, we applied the Data Preprocessing step to the sampled clinical notes. This operation included the extraction of relevant sections from the clinical notes (Section Extraction) and the application of textual preprocessing techniques, such as stopwords elimination and lemmatization, to the text of the clinical notes.

3.4. Summarization

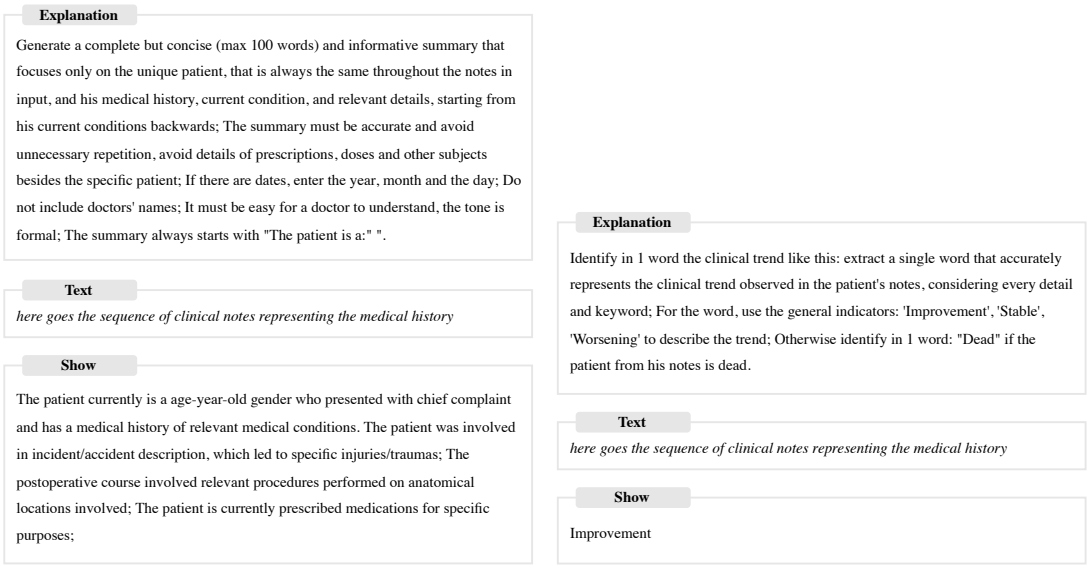
To generate the summaries of the clinical notes, we used a model based on Large Language Models (LLMs). The input for the model was structured according to the following prompt:

- **System**: e.g for summary: "You are a formal medical assistant specialising in the summary of a patient's clinical notes" while for clinical trend: "You are a clinical trend extractor of a patient 's clinical notes."
- **User**: an experimental approach:

$$P = E + T + S \tag{1}$$

  Where:

  - **P**: **P**rompt;
  - **E**: **E**xplanation: explanation of what is wanted from the model;
  - **T**: **T**ext: text that the model must deal with and summarise, in this case the clinical notes;
  - **S**: **S**how the type of the task: demonstration of an example of output useful for the model;



(a) Summary (b) Clinical trend

Figure 2: Examples of summarization prompts.

We applied the same structure for both the general summary of the clinical notes (Summary) and the identification of clinical trends (Clinical Trend).

Summary: the aim of the summary is to provide an overview of the patient's condition, treatments carried out, main diagnoses or other key points in the clinical notes. Figure 2(a) shows an example of summarization prompts.

Clinical trend: refers to patterns or changes observed in a patient's clinical picture over time. Figure 2(b) shows an example of prompt used to infer the medical history trend.

3.5. Named Entity Recognition + Linking + Relation Extraction

From the preprocessed data, we applied the Named Entity Recognition (NER) and Entity Linking (L) process to extract medical entities from the clinical notes and link them to external knowledge databases. In addition, we performed Relation Extraction (RE) to identify possible relationships between entities. This information was used to create a Knowledge Base for graph analysis, in which diagnostic analyses were performed.

3.6. Analytics and Report

We summarized the results obtained from the clinical note summary process and the Named Entity Recognition + Entity Linking + Relation Extraction process. These results were presented in a final report documenting the main results obtained within our study.

4. Results

4.1. Implementation details

The techniques seen above are applied by randomly selecting 20 patients from the MIMIC-III dataset [17]; in particular, only the "Discharge summary" of patients selected from the NOTEEVENTS table is considered[2].

Then, through the **MedSpaCy** [19] module, the extraction of the main sections, i.e., those most present, from all patients' clinical notes is performed. MedSpaCy was chosen because that library is already pre-trained to recognize sections present in clinical texts. Thus, for the case under consideration, the following sections are selected: "chief complaint," "history of present illness," "past medical history," "discharge medications," "brief hospital course" and "discharge diagnoses". In the clinical notes with the extracted sections, a following stage of stopwords elimination and lemmatization is carried out through the **NLTK**[3] module using stopwords of the English language, the clinical notes being in that language, and WordNet as a lemmatizer.

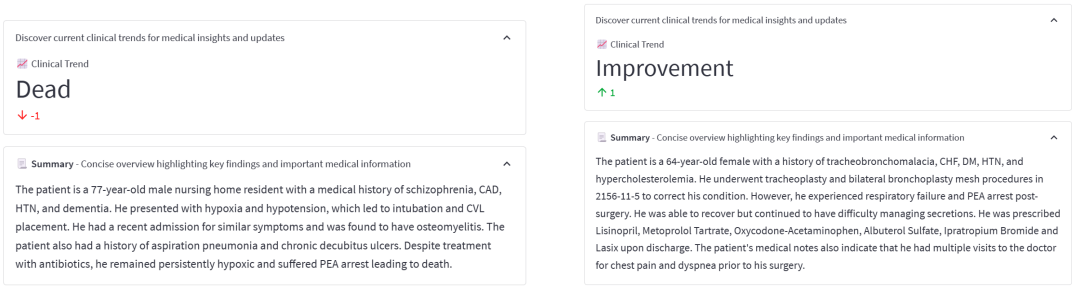
To perform the summarization task we chose to use as LLM the **GPT-3.5-turbo** [4] model provided by OpenAI, this model, in fact, is the one that starting from a prompt and the clinical notes of the incoming patient manages to return a short, but at the same time complete, summary containing all the main information regarding the patient's medical history. In addition to the patient summary, the patient's clinical trend is also generated using the same model, which, based on the evolution of his clinical history shown in the notes, provides an indication of the patient's current status.

For the NER task it was decided to use the **MedCAT** [20] model, this model was found to be the one with the best performance for the requested task being pre-trained for natural language processing in the medical field. In particular, it is very useful for extracting information from Electronic Health Records (EHR) (NER phase) and linking them to biomedical ontologies such as SNOMED-CT and UMLS (Entity Linking phase). Since MedCAT is only a model that correctly extracts and labels entities, it is necessary to provide MedCAT with a knowledge base from which to draw this information. For the case in question, it was decided to use the **MedMentions** [21] library, which contains a corpus of biomedical documents annotated with mentions of entities belonging to UMLS. This corpus contains about 35,000 medical concepts and its MetaCAT model for meta annotations was built on a sample from MIMIC-III. An example of a clinical note annotated with MedCAT using MedMentions is shown in Figure 3.

Finally, the concepts extracted from the NER+L phase must be interconnected through the relationships extracted with the Relation Extraction phase, also in this case these relations are obtained using the GPT-3.5-turbo model. Given their extremely complex nature, these relationships are particularly suitable to be stored via a graph database such as **Neo4J**[5].

[2]The complete code used to carry out the experiments reported in the following article is available in the following GitHub repository: https://github.com/giuseppericcio/HealthcareSummarizationMIMIC
[3]Documentation available at the site: https://www.nltk.org/
[4]More details on models provided by OpenAI: https://platform.openai.com/docs/models/gpt-3-5
[5]Documentation available at the site: https://neo4j.com/docs/

Figure 3: Output example of the NER step on a clinical note (patient 27431).



(a) Patient 27431 (b) Patient 30822

Figure 4: Clinical Trend and Summary generated for several patients

4.2. Dashboard and Analytics

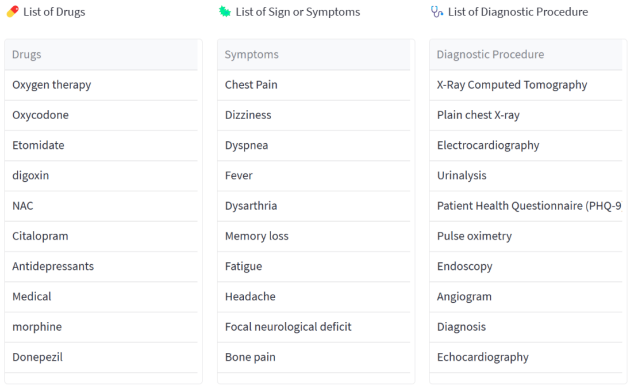
Thanks to the entities extracted from the NER+L+RE phase and the summaries generated by the LLM, it is possible to build a dashboard that supports the medical personnel during the diagnosis and the choice of therapies to be carried out on a patient in order to cure his diseases. In this instance, it was decided to use the **Streamlit**[6] library for the construction of the dashboard directly in Python. This choice is dictated by the simplicity of creating a dashboard using this library which allows various effective views of the analytics that can be performed on biomedical concepts recognized by NER+L+RE. In Figure 4 is possible to see some examples of summaries generated from some clinical notes of MIMIC-III patients, through these summaries a physician can understand all the diseases and procedures with respect to that patient in a fast way than reading all of his clinical notes.

4.2.1. Analytic 1: Lists of concepts extracted

First of all, through the dashboard, it is possible to view lists of drugs, symptoms and diagnostic procedures related to a specific patient, as shown in Figure 5. Through this visualization, medical

[6]Documentation available at the site: https://docs.streamlit.io/

personnel can immediately understand what the patient has been subjected to without having to read all his medical records.



List of Drugs

Drugs
Oxygen therapy
Oxycodone
Etomidate
digoxin
NAC
Citalopram
Antidepressants
Medical
morphine
Donepezil

List of Sign or Symptoms

Symptoms
Chest Pain
Dizziness
Dyspnea
Fever
Dysarthria
Memory loss
Fatigue
Headache
Focal neurological deficit
Bone pain

List of Diagnostic Procedure

Diagnostic Procedure
X-Ray Computed Tomography
Plain chest X-ray
Electrocardiography
Urinalysis
Patient Health Questionnaire (PHQ-9
Pulse oximetry
Endoscopy
Angiogram
Diagnosis
Echocardiography

Figure 5: Lists of concepts extracted by NER+L for the patient 27431

4.2.2. Analytic 2: Diseases of a patient

Using the relations extracted from the NER+L+RE phase is possible to visualize some interesting analytics. Through the concepts stored in Neo4J, which has also been integrated into Streamlit via the py2neo and streamlit-agraph libraries. As shown in Figure 6(a), it is possible to effectively display all the diseases associated with a patient. For example, the patient taken into consideration presented "Hypoxia" in all the medical records associated with him, therefore, it could be deduced that he suffered from it chronically.

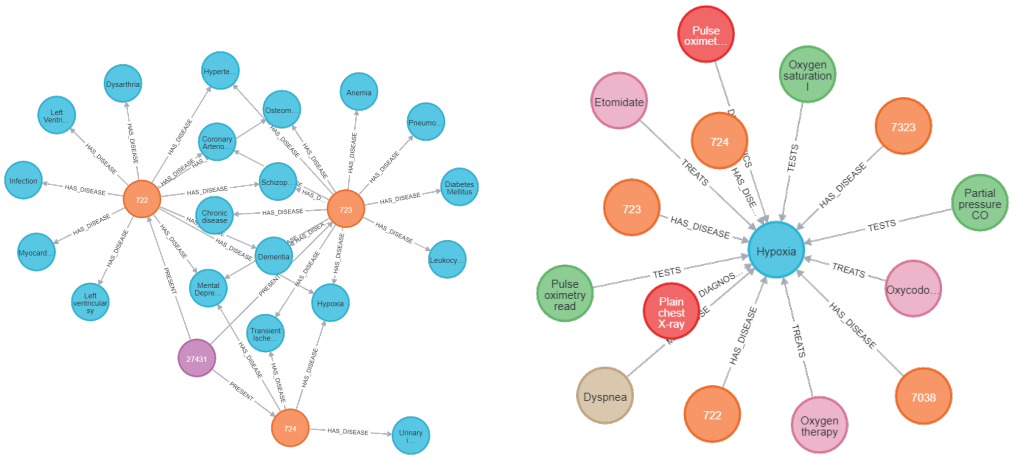
4.2.3. Analytic 3: Medical concepts related to a disease of a patient

With reference to the previous analytic, we can now explore all the diagnostic procedures, drugs and other treatments performed on the patient to cure the "Hypoxia" disease, as shown in Figure 6(b), in order to facilitate physicians and nurses understand what has already been done and what needs to be done later on to that patient.

5. Discussion

Our solution offers numerous benefits, such as automating the clinical note synthesis process, improving productivity and reducing analysis time for healthcare professionals. Through the use of advanced techniques such as NER+L+RE, we are able to identify medical entities and the relationships between them, providing a solid basis for analysing and interpreting clinical information.

However, the automatic extraction and synthesis of information from clinical records presents significant and sensitive challenges. The management of sensitive data, in compliance with regulations such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the Artificial Intelligence Act, requires special



(a) Diseases of the patient

(b) Procedures, drugs and so on related to a specific disease

Figure 6: Graph Analytics for the patient 27431

attention to privacy and the protection of sensitive personal data. Therefore, it is essential to ensure security and compliance with guidelines regarding data access, storage and disposal.

To address these ethical and regulatory issues, robust security measures must be implemented to protect patients' personal data and ensure transparency in the use of clinical information. Furthermore, it is crucial to inform patients about the processing of their data and the generative nature of the results obtained, emphasising that the system does not replace the work of doctors. Adherence to ethical standards is of paramount importance to maintain patients' trust and to ensure responsible and safe use of clinical information. For example, the automatic generation of summaries raises ethical issues regarding the accurate interpretation of clinical information, so it is necessary to ensure the reliability of the generated results and assess the quality of the summaries through further research and validation.

Another disadvantage comes from the tools and limitations of the models used. The specialised medical language used in clinical registries, with abbreviations and technical terminology, requires the use of additional resources, such as medical dictionaries or abbreviation recognition systems, in order to overcome the challenges of information interpretation and extraction. Furthermore, the size of clinical note datasets requires the use of efficient text processing models in order to handle large volumes of data.

Finally, our work represents a step towards automation and optimization of clinical note analysis, but further studies and collaborations are needed to improve the accuracy, reliability and adherence to ethical standards of our solutions.

6. Conclusion and Future Work

Our work has developed a solution for the automated summarization of clinical notes using NER+L+RE and LLM techniques, providing fast decision support for healthcare professionals

towards patients. The preliminary results obtained in this paper were submitted to a committee of domain experts for review, who upon initial analysis evaluated the work positively. However, there are many possibilities for further development and future work arising from this project. Some ideas include:

- **Predictive analytics**: Expanding our solution to include predictive analytics models that can provide estimates of patients' future conditions, such as the likelihood of developing certain diseases or response to certain treatments;
- **Patient profiling**: Create a comprehensive overview of all patients treated, allowing clinicians to identify high-risk patients. This would require unsupervised data analysis, such as using clustering algorithms to group patients according to common characteristics;
- **Interactivity**: Increased interactivity through human body diagrams that display diseased or clinically affected body parts with a brief summary of the problem;
- **Integration**: Expand our solution to be easily integrated with databases from different hospitals, allowing healthcare professionals to use the system with their own data;
- **Interpretability**: Improve the transparency and interpretability of the system by providing clear explanations of the forecasts and recommendations generated. This would help physicians understand the reasons behind the results and have confidence in the information provided.
- **Q/A (Question/Answer)**: Implement a Q/A interface that allows doctors and patients to interact directly with the system, asking specific questions and getting precise answers based on the data in the database.

In conclusion, the future goal is to continue to develop solutions that improve the efficiency (currently the proposed solution on 20 patients takes an average time of 298 seconds) and accuracy of clinical note analysis through systematic and more formal approaches.

References

[1] D. Charles, J. King, V. Patel, M. Furukawa, Adoption of electronic health record systems among u.s. non-federal acute care hospitals, ONC Data Brief No. 9 (2013) 1–9.

[2] D. Soomro, S. Banbhrani, A. Shaikh, H. Raj, Bio-ner: Biomedical named entity recognition using rule-based and statistical learners, 2017.

[3] T. Loftus, B. Shickel, J. Balch, P. Tighe, K. Abbott, B. Fazzone, E. Anderson, J. Rozowsky, T. Ozrazgat Baslanti, Y. Ren, S. Berceli, W. Hogan, P. Efron, J. Moorman, P. Rashidi, G. Upchurch, A. Bihorac, Phenotype clustering in health care: A narrative review for clinicians, Frontiers in Artificial Intelligence 5 (2022) 842306. doi:10.3389/frai.2022.842306.

[4] M. K. Rohil, V. Magotra, An exploratory study of automatic text summarization in biomedical and healthcare domain, Healthcare Analytics 2 (2022) 100058. doi:10.1016/j.health.2022.100058.

[5] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, J. Tsujii, brat: a web-based tool for nlp-assisted text annotation, The 3th Conference of the European Chapter of the Association for Computational Linguistics; Avignon, France (2012) 102–107.

[6] A. Aronson, Effective mapping of biomedical text to the umls metathesaurus: The metamap program, Proceedings / AMIA ... Annual Symposium. AMIA Symposium 2001 (2001) 17–21.
[7] G. Savova, J. Masanz, P. Ogren, J. Zheng, S. Sohn, K. Kipper-Schuler, C. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications, Journal of the American Medical Informatics Association : JAMIA 17 (2010) 507–13. doi:10.1136/jamia.2009.001560.
[8] H. Wu, G. Toti, K. Morley, Z. Ibrahim, A. Folarin, R. Jackson, I. Kartoglu, A. Agrawal, C. Stringer, D. Gale, G. Gorrell, A. Roberts, M. Broadbent, R. Stewart, R. Dobson, 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 25 (2018) 160. doi:10.1093/jamia/ocx160.
[9] J. Cheng, M. Lapata, Neural summarization by extracting sentences and words, 2016. doi:10.18653/v1/P16-1046.
[10] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, Proceedings of the AAAI Conference on Artificial Intelligence 31 (2016). doi:10.1609/aaai.v31i1.10958.
[11] E. Egonmwan, Y. Chali, Transformer-based model for single documents neural summarization, 2019. doi:10.18653/v1/D19-5607.
[12] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 2019.
[13] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mass: Masked sequence to sequence pre-training for language generation, 2019. arXiv:1905.02450.
[14] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H.-W. Hon, Unified language model pre-training for natural language understanding and generation, 2019. arXiv:1905.03197.
[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. arXiv:1910.13461.
[16] A. Johnson, T. Pollard, M. Roger, "mimic-iii clinical database" (version 1.4), 2016. doi:10.13026/C2XW26.
[17] A. Johnson, T. Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 160035. doi:10.1038/sdata.2016.35.
[18] A. Goldberger, L. Amaral, L. Glass, S. Havlin, J. Hausdorg, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng, H. Stanley, P. Physiobank, Components of a new research resource for complex physiologic signals, PhysioNet 101 (2000).
[19] H. Eyre, A. Chapman, K. Peterson, J. Shi, P. Alba, M. Jones, T. Box, S. DuVall, O. Patterson, Launching into clinical space with medspacy: a new clinical text processing toolkit in python, AMIA ... Annual Symposium proceedings. AMIA Symposium 2021 (2022) 438–447.
[20] Z. Kraljevic, D. Bean, A. Mascio, L. Roguski, A. Folarin, A. Roberts, R. Bendayan, R. Dobson, Medcat – medical concept annotation tool, 2019. arXiv:1912.10166.
[21] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with umls concepts, 2019.