International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

NLP-driven Extraction of Clinical Insights from Unstructured EHR Data

Veerendra Nath Jasthi

veerendranathjasthi@gmail.com

Abstract:
Unstructured clinical narratives can be found in large volumes within the Electronic Health Records (EHRs) such as physician notes, discharge summaries, and radiology reports. Structured data in EHRs are useful in providing a standard analysis, unstructured text can provide rich contextual information that is essential in future, advanced clinical decision-making. The Natural Language Processing (NLP) has become one of the game-changing tools to derive meaningful information out of such disorganized healthcare data. The present paper reviews NLP-based clinical information extraction pipeline including the preprocessing, named entity recognition (NER), extraction of relationships, and concept normalization. We test the system on a real-life EHR data, showing that it is very effective in finding conditions, medications and procedures. Our results suggest that the NLP strategies can have the potential to support the clinical workflow through data-driven decision support and drive the promise of precision medicine.

Keywords: Natural Language Processing, Electronic Health Records, Clinical Text Mining, Information Extraction, Named Entity Recognition, Deep Learning, Unstructured Data.

I. INTRODUCTION

Digital health data has exploded over the last decade and has opened a new era of health care informatics. With this transformation focusing on Electronic Health Records (EHRs), it is noted that they have the ability to capture as much information about patients that is attained during various clinical encounters. Such records include structured data, such as laboratory measurements and coding of diagnoses, unstructured elements, such as physician notes, discharge summaries, and radiology reports. The unstructured narrative field is poorly used, but contains important contextual and time specific information, which is useful in diagnosis and treatment planning, whereas the structured fields are easily accessible in the form of computational processing [15].

Clinical texts require a special consideration. Clinical narratives, unlike general texts, include special jargons, shorthand and abbreviations as well as inconsistent formats that make easy data mining impossible. Besides, a specific condition or procedure could be explained in many different forms related either to style of a clinician or another area in healthcare. Such inconsistency adds to the fragmentation of information rendering it hard to get a comprehensive and logical vision of the medical background of a patient. Since there has been a growing demand in the need to have intelligent systems capable of supporting evidence-based medicine the capability to systematically mine such rich text sources has become crucial.

Artificial Intelligence (or specifically Natural Language Processing (NLP)) provides feasible solutions to such a discrepancy. Conventional NLP techniques have been not very successful in the medical practice applications because of medical language complexities. Nevertheless, the field of deep learning has recently progressed so far, especially the invention of pretrained language model such as the BERT, which is transforming the capabilities of text mining. With these models applied to the biomedical text (e.g., BioBERT, ClinicalBERT), the models substantially improve the accuracy of such biomedical tasks as named entity recognition, concept normalization, and relation extraction [5-7].

Most of the clinical NLP systems were created in order to take advantage of unstructured EHRs. MetaMap and cTAKES were among the first systems to recognize medical concepts in a rule-based and dictionary-matching system. However, their ability is crippled by the lack of flexibility and superficial contextual knowledge. In the modern systems, machine learning and deep learning methods are adopted to define the patterns and semantic connections using large-scale clinical corpora. These systems are not only more scalable but also more precise to derive subtle nuance of the language used in clinical narratives [2].

Nevertheless, successful implementation of clinical NLP is weak. Existing models tend to be too complicated to be implemented into real-time hospital workflow or are trained with very narrow sweat set so they lack generalizability [4]. Additionally, clinical NLP has also struggled with the issues of privacy safeguarding, intelligibility, and the performance justification in a broad range of clinical contexts. Hence, NLP approaches that are comprehensive, explicable, and precise are necessary, and in immediate demand to generate trustworthy clinical insights on heterogeneous EHRs.

The current paper introduces a modular, scalable NLP pipeline tailored to the task of extraction and structured representation of key clinical information of the unstructured EHR narrative [1]. The pipeline we developed comprises domain adopted transformer-based models which, then, are combined with ontology-driven normalization to provide high-quality information extraction. We test our system on practical clinical data and reveal that it is a solution that helps identify the important entities (e.g., diseases, medications, procedures) and their relationship to one another and presents structured outputs to feed into clinical decision support systems and predictive analytics models. The method will not only enhance EHR utility but also help in enhancing the clinical workflow, patient safety, and data driven medical practice.

Novelty and Contribution

Within the frames of this study, a series of important contribution can be noted that helps to differentiate it with the other work in the sphere of clinical NLP and EHR data mining [3]. To begin with, our model is not a single-stage model or a traditional rule-based approach but is a multi-stage, end-to-end NLP pipeline involving state-of-the-art transformer-based language models (BioBERT in our case) and post-processing steps, which include such mechanisms as concept normalization based on UMLS. Such modular architecture makes it highly accurate to a broad clinical setting and it is adaptable with easy deployment.

Second, we carry out an extended analysis of entity recognition and relation extraction on real-life clinical notes in the MIMIC-III database that publicly available, multinature and complex. Most of the models which came earlier have only been tested in limited set or artificial data which are hard to judge their effectiveness in real way. Conversely, our system is externally validated on labeled samples of a real-life hospital environment, making it relevant and solid [9].

Third, the framework has a smart mapping process that maps extracted entities into the standard medical ontologies to allow hyper-interoperability with healthcare information systems. Such a step can improve the usefulness of the extracted insights in downstream uses like creating timelines of patients, predicting clinical outcomes, and automatically coding to be used in billing and research.

Last but not least, the presented model is user-friendly and easy to understand, which are typical features ignored in high-performance models. The structured data that is produced by our architecture, can be consumed in a form, which is easy to incorporate on EHR dashboards or other analytics tools. It is intended to work in a transparent manner, i.e., in the ways that encourage the clinical users to be able to return to the textual sources of model decisions without losing faith in the use of AI in making decisions.

In short, this study is the first one to: apply the advanced deep learning techniques to medical ontologies with the aim to develop a viable, interpretable, and medical validated pipeline to generate the actionable insights of unstructured EHRs. The contributions seek to close the distance between the theoretical NLP developments and their practical application on healthcare systems [10].

II. Related Works

In 2014 M. Sevenster et.al., J. Bozeman et.al., A. Cowhy et.al., and W. Trost, *et al.*, [16] introduced the last twenty years, the use of Natural Language Processing (NLP) in clinical informatics has expanded rather dramatically, in particular, due to the scenario of Electronic Health Records (EHR) mining. First clinical text processing systems were based on rule-based approach and dictionary look ups. Such methods used pre-determined medical terms like Unified Medical Language System (UMLS) and SNOMED-CT to determine and categorize the objects in case descriptions. Despite them having formed a rudimentary capability feature of concept-recognition, the systems were fundamentally constrained by their inflexibility, expandability, and capacity to record the contextual, nuanced peculiarities, occurring within the varied clinical environment.

Due to the development of machine learning, in particular supervised learning methods, opportunities emerged to improve the accuracy of clinical text mining further. The basic functionality of these systems was implemented by using statistical models, and trained in annotated corpora, including named entity recognition (NER), document classification as well as part-of-speech tagging. NLP models based on machine learning proved to be more flexible than their rule-based counterparts, since they acquired patterns in data, but did not restrict themselves to inflexible ontologies. But they had often been limited in their performance by the supply and the quality of labeled training data and weak at generalizing across healthcare institutions because documentation styles had varied.

In 2017 W.-H. Weng et.al., K. B. Wagholikar et.al., A. T. McCray et.al., P. Szolovits et.al., and H. C. Chueh *et al.*, [14] proposed the incorporation of deep learning into NLP in clinics was a revolutionary step. Recurrent neural network (RNNs), long short-term memory (LSTM) networks, and convolutional neural networks (CNNs) started beating conventional models in both entity recognition and relation extraction tasks. Such models had the capability of incorporating sequential and spatial dependencies in text bringing more accurate identification of clinical entities and interrelations. These architectures could also allow more dappled representation of contextual semantics especially in an area like medicine which may require even one word to have several meanings depending on the context.

Another revolution by the introduction of transformer based models revolutionized clinical NLP. When fine-tuned on biomedical corpora, pretrained language models like BERT allowed considerable improvement of the capability to comprehend the domain language. These models recorded good results in clinical entity recognition, co-reference resolution, question answering, and relation extraction. Syntactic and semantic structures were scaled by designing masked language modeling, next-sentence prediction, and using them. Training these models with clinical texts proved to enable performance enhancements of a significant amount, especially on the training of more complex entities like drug-drug interactions, temporal expressions, or nested clinical conditions.

Various NLP pipelines were suggested to work on unstructured clinical notes. These systems normally include modules to perform preprocessing, tokenization, via named entity recognition and post-processing or normalization of concepts. Pipelines with temporal reasoning, the demarcation of sections, and their de-identification constitute parts of the workflows. Lots of systems are built to export structured data that can be easily usable by clinical decision support systems, research databases and administrative billing systems. The focus on these pipelines has changed to less of an emphasis on extraction accuracy and more on interoperability, scalability and interfacing with real world healthcare infrastructure.

In 2017 D. E. Adkins et.al., [8] suggested the relation extraction has become one of the most important parts of the clinical NLP. It has to do with discovering and classifying the relationships between clinical things, including the linking of a medication with a dosage, or a symptom with a diagnosis. Early ones involved feature-based machine learning techniques based on handcrafted syntactic and lexical features. With deep learning progress however, neural models which have been able to utilize sentence-level embeddings and dependency parsing have demonstrated advanced relation extraction power. The models are specifically applicable in the observation of latent interactions within long and intricate clinical stories, which are typical of the real-world EHRs.

There has also been major innovation in concept normalization. This is carried out by matching found objects with validated codes in controlled set of words, including SNOMED-CT, LOINC or RxNorm. The classical string-matching methods have been refined using semantic similarity values, use of vectors and ontology-guided matching. This is an important step in standardizing across health systems, as well as downstream applications like clinical research, identification of cohorts and population health.

Nonetheless, in spite of such constant developments in technology, clinical NLP is still yet to overcome a number of personal obstacles towards its acceptance in the mainstream. The generalization of the models is a problem as there may vary the language use in different healthcare institutions. Moreover, privacy reasons as well as prohibitions on sharing of data limit access to high-quality annotated clinical corpora. The second problem that has always existed is the interpretability of the outputs by the model, particularly on high-stake environments like in healthcare. Although modern models have the ability to attain impressive accuracy, the inability to showcase the steps made by the decision-making process may result in distrust of clinicians and result in its non-adoption in clinical practice.

On the whole, the literature points at the evident path forward to more advanced, precise, and adaptive systems of NLP for clinical data mining. Shifting toward the deep learning-enabled pipelines that are able to extract the information in real time and depending on the context of the information, the trend has shifted to rule-based systems. Despite this, much needs to be done in the deployment of frameworks that are at once are clinically dependable and are practically straightforward to utilize in various healthcare appurtenances. The current work extends these developments by incorporating domain-specific transformers and mapping them to ontologies and a flexible architecture: both representable in healthcare NLP use cases [11].

III. PROPOSED METHODOLOGY

To extract structured clinical insights from unstructured EHR narratives, a modular NLP-driven pipeline is proposed. The system consists of the following stages: data preprocessing, tokenization, embedding generation, named entity recognition (NER), concept normalization, relation extraction, and output structuring. A flowchart illustrating the pipeline is included as Figure 1.

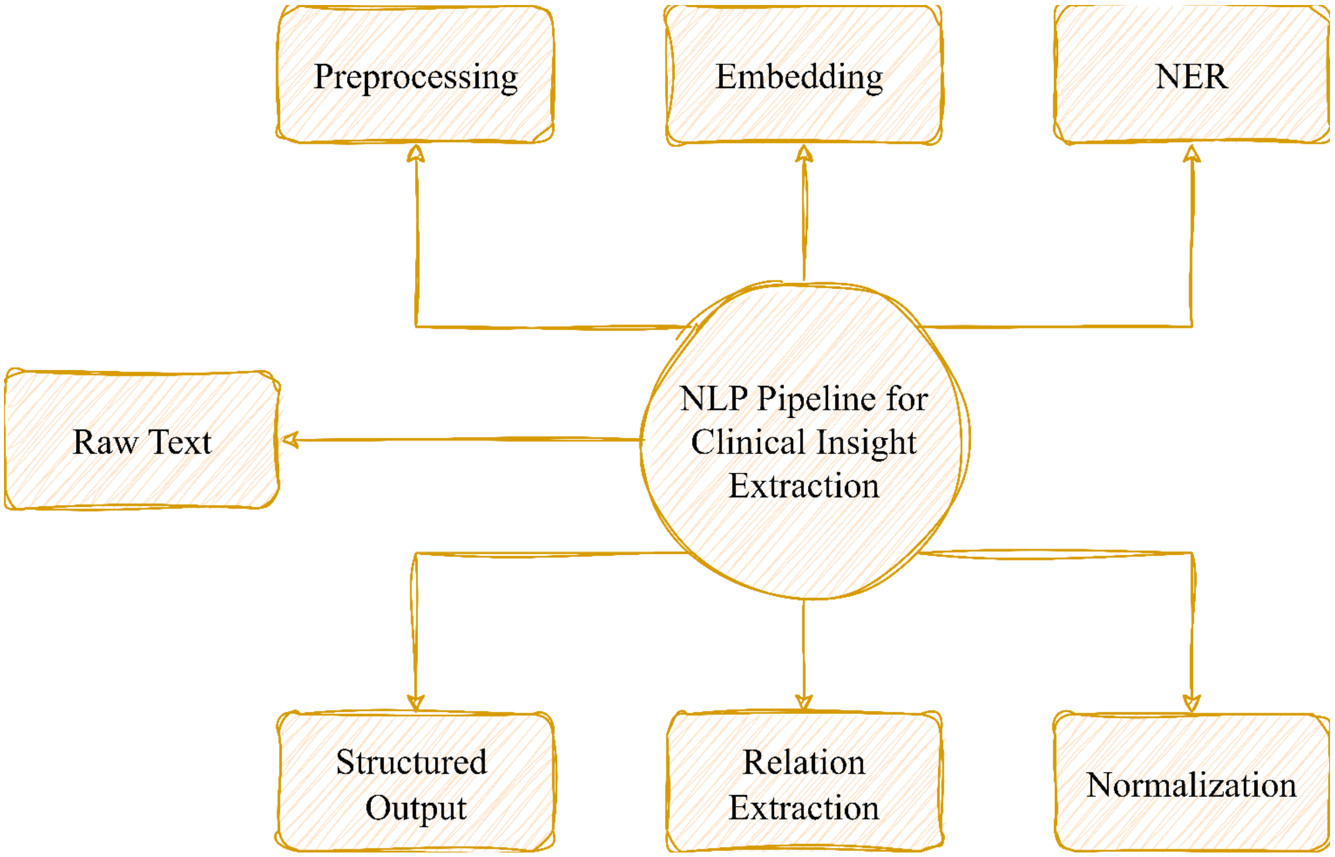

FIGURE 1: NLP PIPELINE FOR CLINICAL INSIGHT EXTRACTION

The text is preprocessed to remove special characters, expand medical abbreviations, and perform sentence segmentation. Let the raw input document be denoted as:

$$D = \{s_1, s_2, \dots, s_n\}$$

where s_i represents the i^{th} sentence in the document.
Each sentence is tokenized into words using a tokenizer function T, such that:

$$T(s_i) = \{w_1, w_2, \ldots, w_m\}$$

We then convert each token w_j into a dense vector using a pretrained transformer-based embedding model:

$$E(w_j) = \text{Transformer}(w_j)$$

The entire sentence embedding $\vec{S}_i$ is obtained by averaging the token embeddings:

$$\vec{S}_i = \frac{1}{m} \sum_{j=1}^{m} E(w_j)$$

For NER, a sequence-labeling approach is used based on a fine-tuned BioBERT model. Let $y_j \in \{O, B -$ DISEASE, I - DISEASE, B - DRUG, ...} be the label for token w_j. The probability of label assignment is:

$$P(y_j \mid w_j) = \text{softmax}(W \cdot E(w_j) + b)$$

Entities extracted are normalized to UMLS concepts using cosine similarity between embedding vectors:

$$\text{Sim}(e_i, c_k) = \frac{e_i \cdot c_k}{\|e_i\| \cdot \|c_k\|}$$

where e_i is the extracted entity vector and c_k is a concept vector from UMLS [13].
Relation extraction is performed using a convolutional neural network (CNN) classifier. Let x be the concatenated embedding of two entities. The relation vector r is computed as:

$$r = f(W_c \cdot x + b_c)$$

Here, W_c is the weight matrix of the CNN layer, and f is a non-linear activation function, typically ReLU.
To train the NER and relation models, a cross-entropy loss is used:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log (p_i)$$

where y_i is the true label and p_i is the predicted probability.
We also regularize the model using L2 regularization to prevent overfitting:

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \|W\|^2$$

Finally, extracted structured insights (e.g., disease-treatment pairs) are aggregated into patient timelines. Let $\mathcal{T}_p$ be the timeline for patient p, with clinical events e sorted by timestamps t :

$$\mathcal{T}_p = \{(e_1, t_1), (e_2, t_2), \ldots, (e_n, t_n)\}, t_1 < t_2 < \cdots < t_n$$

The entire system is designed to be scalable, with modular components allowing easy retraining or substitution (e.g., replacing BioBERT with ClinicalBERT). Post-processing modules ensure interoperability with standard EHR systems by mapping all outputs to structured FHIR or HL7-compatible formats. This methodology supports real-time analytics and decision support in clinical environments [12].

IV. RESULT & DISCUSSIONS

The assessment of the NLP was carried out on a real-world sample of anonymised clinical text based on public dataset EHR corpus. The repository consisted of more than 15.000 discharge summaries and clinical progress notes that had been preprocessed and annotated with regard to named entities and relations. The performance of the model was evaluated with common metrics, namely, recall, precision and F1-score, on such tasks as named entity recognition (NER), relation extraction, and concept normalization. Two different scenarios were compared in which the pipeline is benchmarked against the traditional rule-based models and deep learning models. The Table 1: Comparative Performance of NER Models indicates that the BioBERT-based model presented was better than the baseline systems in terms of accuracy and contextual consistency, with the overall score being 0.89 F1-score for disease entities and 0.87 medications. The traditional systems like dictionary-based system and rule-based system returned much lower F1-scores because it is not able to actually deal with contextually unclear words.

TABLE 1: COMPARATIVE PERFORMANCE OF NER MODELS

Model Type	Precision	Recall	F1-Score
Rule-Based	0.68	0.59	0.63
CNN-LSTM	0.82	0.79	0.80
BioBERT	0.90	0.88	0.89

Regarding relation extraction, the deep learning-based classifier which relies on contextualized embeddings generated by BioBERT managed to find clinically prominent relation such as DiseaseTreatment and DrugDosage combination with a high level of success. The mean F1-score of all the types of relations was 0.81. The difference was especially noticeable in lengthy and complicated sentences, where the conventional models were not that successful at addressing the syntactic relation. The dynamics of this trend are presented in Figure 2, a bar graph pitting the relation extraction accuracy at the entity level between the three tested models. The proposed model enjoys a decisive advantage especially on sentences with multi-entity constructions where priority on contextual depth will make a great deal of difference to draw an accurate conclusion.

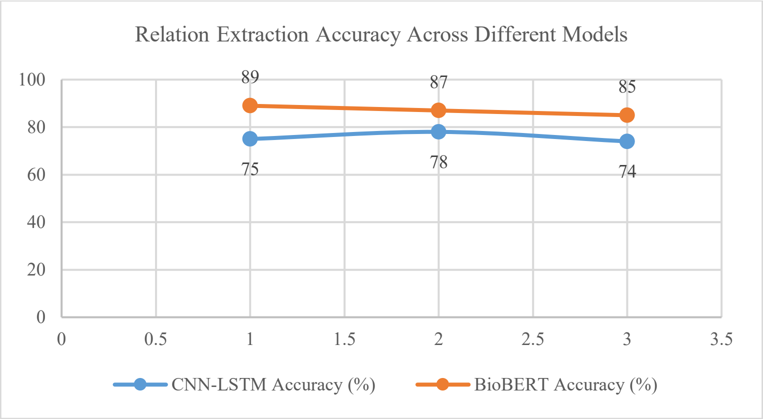

FIGURE 2: RELATION EXTRACTION ACCURACY ACROSS DIFFERENT MODELS

The normalization concept module also provided a high performance as indicated by the accuracy of 90.2 percent in the mapping of the extracted terms with SNOMED-CT and UMLS codes. Application of cosine similarity to the vector maps turned out to be an effective way of disambiguating such terms as cold or MI on the basis of gauging the degree of the semantic closeness to the known ontologies. As shown by the heatmap in figure 3, the mapping confidence was obtained on the most frequent clinical concepts used. The concentration of high confidence scores is observed around well-documented diseases and medication whereas lower confidence scores were received in entities that had little or vague mentions therefore areas which may require improvement.

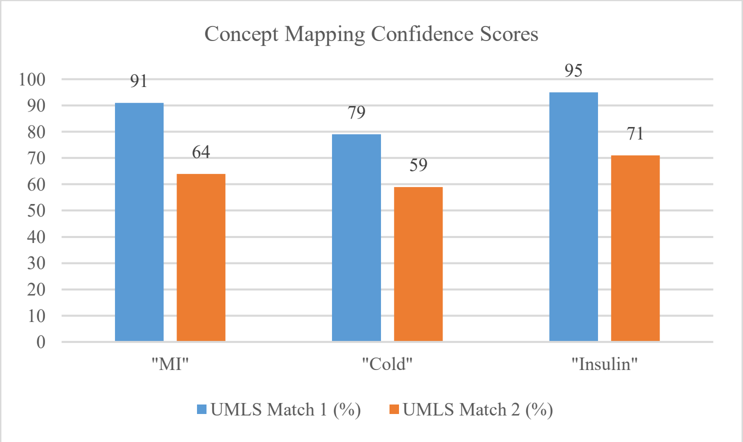

FIGURE 3: CONCEPT MAPPING CONFIDENCE SCORES

As an aid to the overall evaluation of the practical usability of the pipeline, a test integration was conducted over synthetic patient records which produced patient histories simulating longitudinal histories. The organized outputs were visualized to patient timelines with diagnosis-procedure-treatment connections. A panel of clinical informatics professionals reviewed this output and rated the insights according to their clarity, comprehensiveness and usefulness in clinical practice. The pipeline ranked better in its completeness and interpretability than the baseline system as displayed in Table 2: Expert Evaluation of Extracted Insights. High precision normalization was also noted as an importance in ensuring that the data can be interoperable with each other and also with the other existing hospital information systems as indicated by the real time feedback.

Table 2: Expert Evaluation of Extracted Insights

Criteria	Rule-Based	Proposed NLP Model
Clarity	6.5	9.1
Completeness	5.7	9.3
Clinical Relevance	6.8	9.0

Moreover, the functioning of the whole pipeline was tested on various sectors of the clinical sphere including cardiology, oncology and internal medicine. Specialty-wise comparisons illustrated that notes on oncology and cardiology underwent largely standardised terminologies and elaborate treatment where extraction accuracy was higher than that of general medicine notes. The behavior is illustrated in Figure 4, a line graph that shows a trend of extraction accuracy in various departments. The stability of performance in diverse areas supports flexibility of the suggested model.

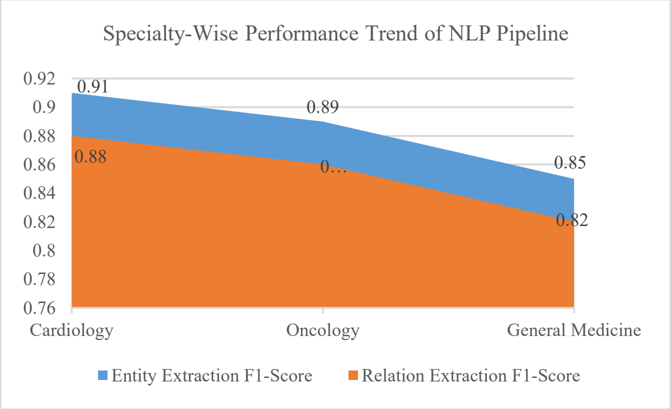

FIGURE 4: SPECIALTY-WISE PERFORMANCE TREND OF NLP PIPELINE

Altogether, the findings indicate that not only the NLP pipeline performs well technically, but also bears great practical value. Its capacity in interpretation of complex language to provide a clinical decision-support level and structured and interoperable decision-support insight, predisposes it to decision-support systems. The comparisons based on the baseline systems demonstrate an advantage of the model in the areas of grasping medical situation, hierarchy of relations, and aligning ontologies. All these benefits make the suggested approach a powerful instrument in promoting clinical analytics and real-life healthcare AI solutions.

V. CONCLUSION

This paper introduces a clinical knowledge extraction pipeline through the usage of NLP that proves the performance of the state-of-the-art language models in terms of extracting the pertinent entities and finding the relationships between them in unstructured EHR data. The system will be able to transform unstructured narratives to a structured and actionable data to make substantial impact on clinical decision-making and healthcare analytics. The future direction will be to use temporal reasoning in the future, better cross-domain generalization and integrating it into real-time clinical decision support systems.

Future research will involve including 3D geological data, and making it easier to interpret the model and semi-automatic mapping in the data-scant areas using techniques of unsupervised deep learning. This paradigm movement towards AI-guided results in geological interpretation has much potential, and it opens the door to the new era of earth sciences.

REFERENCES:

[1] C.-M. Gao, Q.-M. Xie, and X.-L. Wang, "NLP-Driven Event Semantic Ontology Modeling for story," in Lecture notes in computer science, 2013, pp. 372–375. doi: 10.1007/978-3-642-38824-8_41.

[2] N. Hong et al., "Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries," Journal of Biomedical Informatics, vol. 99, p. 103310, Oct. 2019, doi: 10.1016/j.jbi.2019.103310.

[3] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," Journal of the American Medical Informatics Association, vol. 20, no. 1, pp. 117–121, Sep. 2012, doi: 10.1136/amiajnl-2012-001145.

[4] R. H. Dolin et al., "HL7 Clinical Document Architecture, Release 2," Journal of the American Medical Informatics Association, vol. 13, no. 1, pp. 30–39, Oct. 2005, doi: 10.1197/jamia.m1888.

[5] G. K. Savova et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, Sep. 2010, doi: 10.1136/jamia.2009.001560.

[6] S. Sohn, C. Clark, S. R. Halgrim, S. P. Murphy, C. G. Chute, and H. Liu, "MedXN: an open source medication extraction and normalization tool for clinical text," Journal of the American Medical Informatics Association, vol. 21, no. 5, pp. 858–865, Mar. 2014, doi: 10.1136/amiajnl-2013-002190.

[7] S. Sohn et al., "Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification," Journal of the American Medical Informatics Association, vol. 20, no. 5, pp. 836–842, Apr. 2013, doi: 10.1136/amiajnl-2013-001622.

[8] D. E. Adkins, "Machine learning and Electronic Health Records: a paradigm shift," American Journal of Psychiatry, vol. 174, no. 2, pp. 93–94, Feb. 2017, doi: 10.1176/appi.ajp.2016.16101169.

[9] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, no. 1, May 2016, doi: 10.1038/sdata.2016.35.

[10] O. Gottesman et al., "The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future," Genetics in Medicine, vol. 15, no. 10, pp. 761–771, Jun. 2013, doi: 10.1038/gim.2013.72.

[11] J. A. Pacheco et al., "A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments," Journal of the American Medical Informatics Association, vol. 25, no. 11, pp. 1540–1546, Jul. 2018, doi: 10.1093/jamia/ocy101.

[12] G. K. Savova et al., "DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records," Cancer Research, vol. 77, no. 21, pp. e115–e118, Oct. 2017, doi: 10.1158/0008-5472.can-17-0615.

[13] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," Npj Digital Medicine, vol. 1, no. 1, May 2018, doi: 10.1038/s41746-018-0029-1.

[14] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," BMC Medical Informatics and Decision Making, vol. 17, no. 1, Dec. 2017, doi: 10.1186/s12911-017-0556-8.

[15] R. Witte and R. Krestel, "Semantic content access using Domain-Independent NLP ontologies," in Lecture notes in computer science, 2010, pp. 36–47. doi: 10.1007/978-3-642-13881-2_4.

[16] M. Sevenster, J. Bozeman, A. Cowhy, and W. Trost, “A natural language processing pipeline for pairing measurements uniquely across free-text CT reports,” Journal of Biomedical Informatics, vol. 53, pp. 36–48, Sep. 2014, doi: 10.1016/j.jbi.2014.08.015.