Leveraging Clinical BERT for End-to-End Knowledge Discovery and Graph Representation in Healthcare: A Comprehensive Approach

Naveen S. Pagad
Assistant Professor, Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India

Pradeep N
Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India.

E-mail (CorrespondingAuthor): naveenspagad@gmail.com
Contact Number(CorrespondingAuthor): 9964006531

Abstract:

In the ever-evolving landscape of healthcare, the utilization of natural language processing (NLP) techniques has emerged as a pivotal tool for uncovering valuable insights from clinical text data. In this paper, we present a novel approach that leverages the power of Clinical BERT, an advanced NLP model fine-tuned for medical text, to enable end-to-end knowledge discovery and graph representation in healthcare. Our methodology encompasses a comprehensive process that begins with the collection and pre-processing of clinical text data, followed by fine-tuning of the Clinical BERT model to extract entities and relationships. Subsequently, we employ NetworkX, a Python library for graph analysis, in conjunction with Matplotlib for visualization, to construct and analyze a Knowledge Graph representation of the extracted information. Through this integrated approach, we demonstrate the capability to uncover intricate medical insights, identify meaningful relationships between medical concepts, and represent them in a structured and interpretable manner. Our findings showcase the potential of leveraging state-of-the-art NLP techniques in conjunction with graph-based representations to advance healthcare research, clinical decision-making and patient outcomes.

Keywords: Clinical Bert, NetworkX, Knowledge Graph and NLP.

1. Introduction

The intersection of natural language processing (NLP) and healthcare [1] represents a frontier of innovation, offering transformative opportunities to leverage textual data for enhanced clinical decision-making and patient care. At the forefront of this convergence is Clinical BERT, an advanced variant of the BERT model meticulously tailored to the intricacies of medical language. Clinical BERT's capability to comprehend the nuances of clinical text and extract meaningful insights has positioned it as a pivotal tool in revolutionizing healthcare practices [2]. In this section, we embark on an in-depth exploration of Clinical BERT's role in facilitating knowledge discovery and representation within the healthcare landscape.

Clinical BERT, with its foundation in deep learning and Transformer architecture, exhibits a remarkable aptitude for processing clinical text data. By leveraging bidirectional attention mechanisms, Clinical BERT effectively captures contextual relationships between medical terms, enabling accurate identification of entities such as diseases, symptoms, medications, and procedures [3]. This nuanced understanding of medical language forms the cornerstone for subsequent tasks, including named entity recognition (NER) and relationship extraction (RE), which are essential for knowledge discovery.

The knowledge discovery process enabled by Clinical BERT encompasses a multifaceted approach, commencing with the aggregation and preprocessing of diverse clinical text sources. Leveraging its fine-tuned capabilities, Clinical BERT extracts clinically relevant entities and infers semantic relationships between them. This process culminates in the construction of a structured representation of medical knowledge, typically in the form of a Knowledge Graph [4]. The Knowledge Graph serves as a comprehensive repository of medical concepts and their interconnections, facilitating intuitive exploration and interpretation of complex clinical data.

Finally, the integration of Clinical BERT into healthcare workflows heralds a paradigm shift in how medical knowledge is harnessed and utilized. By empowering clinicians and researchers with actionable insights extracted from vast repositories of clinical text data, Clinical BERT holds the promise of enhancing diagnostic accuracy, improving treatment efficacy, and ultimately, advancing patient outcomes. Through our exploration of Clinical BERT's capabilities in knowledge discovery, we aim to underscore its transformative potential in reshaping the landscape of healthcare delivery and catalyzing innovation in medical practice.

2. Literature Survey

In addition to named entity recognition (NER), relationship extraction (RE), and knowledge graph construction, several other aspects of natural language processing (NLP) in the clinical domain have garnered significant attention from researchers. One such area is clinical text classification, which involves categorizing clinical documents or narratives into predefined categories or classes. For example, Hassanpour et al. (2016) [5] explored the use of machine learning algorithms for classifying radiology reports into diagnostic categories, demonstrating the potential of NLP-based approaches in automating document classification tasks in healthcare settings.

Many studies have been conducted with the goal of using NLP methods to extract meaningful information from clinical text data. Entity recognition (NER) is a prominent area of focus that entails recognising and categorising items in clinical narratives, including illnesses, symptoms, drugs, and treatments. The difficulties associated with clinical NER have been addressed by a number of strategies, including machine learning algorithms, statistical models, and rule-based techniques. To increase the accuracy of entity recognition in clinical text, Huang et al. (2019) [6] developed a hybrid strategy that blends rule-based patterns with machine learning techniques.

The advent of pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) has revolutionized the field of clinical NLP by providing powerful representations of textual data. Fine-tuning pre-trained BERT models on clinical text data has been shown to enhance their performance on various clinical NLP tasks, including NER, RE, and knowledge graph construction. Alsentzer et al. (2019) [7], for example, achieved the latest developments on standard datasets by demonstrating the efficacy of fine-tuned BERT frameworks for identifying clinical entities and connections from electronic health records.

In addition to NER, researchers have explored the application of relationship extraction (RE) techniques to uncover meaningful associations between medical concepts mentioned in clinical text. RE tasks aim to identify semantic relationships such as treatment relationships, co-occurrences, and causality between entities. Zhou et al. (2020) [8] proposed a deep learning-based approach for extracting drug-drug interactions from electronic health records, demonstrating the efficacy of neural network models in capturing complex relationships within clinical text data.

Furthermore, the construction of knowledge graphs from clinical text data has emerged as a promising approach for representing and organizing medical knowledge in a structured format. Knowledge graphs enable intuitive exploration of relationships between medical concepts and facilitate advanced analytics and decision support systems. A recent study by Zhang et al. (2021) [9] employed graph-based techniques to construct a knowledge graph from electronic health records, demonstrating its utility in supporting clinical decision-making and biomedical research.

Overall, the literature survey highlights the breadth and depth of research efforts in the field of clinical NLP, spanning a wide range of tasks and applications. Researchers are continually pushing the limits of what NLP can achieve in healthcare, tackling tasks such as named entity recognition, relationship extraction, text classification, sentiment analysis, temporal information extraction, and multimodal data integration. These innovations are poised to enhance healthcare delivery, improve patient outcomes, and propel medical research forward in the coming years.

3. Dataset Collection and Description

The n2cc2 dataset [10, 11] consists of discharge summaries from clinical records, providing a detailed narrative of each patient's medical history, diagnosis, treatment, and discharge condition. Each discharge summary includes key information such as admission and discharge dates, patient demographics (e.g., age, sex), chief complaint, major surgical

procedures, history of present illness, past medical history, social history, family history, physical exam findings, pertinent lab results, imaging reports, electrocardiogram (ECG) findings, EEG results, lumbar spine examination, hospital course, medications on admission and discharge, discharge diagnosis, condition at discharge, and discharge instructions.

The discharge summaries are structured in a consistent format, allowing for easy extraction of information for analysis and processing. They provide a comprehensive overview of the patient's hospitalization, including any complications, treatments administered, and follow-up recommendations. This structured format facilitates the development of natural language processing (NLP) models for tasks such as information extraction, summarization, and predictive analytics.

The dataset is valuable for training and evaluating NLP models in the clinical domain, particularly for tasks such as NER, relationship identification, medical coding, and clinical decision support. By analyzing the text data within these discharge summaries, researchers and healthcare professionals can gain insights into patterns of disease, treatment effectiveness, patient outcomes, and healthcare utilization.

The dataset collection process likely involved anonymizing patient information to ensure privacy and compliance with healthcare regulations. It may have been obtained from electronic health record (EHR) systems, where discharge summaries are routinely generated as part of the documentation process. Additionally, the dataset may have undergone pre-processing steps such as tokenization, sentence segmentation, and cleaning to ensure data quality and consistency. The Fig1 shows a sample record in dataset.

3.1 Features

The features in the dataset refer to the different pieces of information present in each discharge summary that are relevant for analysis and understanding of the patient's medical history, condition, and treatment. Here are some of the key features typically found in such datasets:

1. **Admission Date and Discharge Date:** These features provide the temporal context of the patient's hospitalization, indicating when they were admitted to the hospital and when they were discharged.
2. **Demographic Information**: This includes features such as Date of Birth, Sex, and sometimes additional demographic details like ethnicity or race.
3. **Chief Complaint**: Describes the primary reason for the patient's admission to the hospital, often in their own words or as documented by the healthcare provider.
4. **Medical History**: This includes past medical conditions (e.g., asthma, hypertension), surgical procedures, and any relevant comorbidities.
5. **History of Present Illness (HPI**): Provides a detailed narrative of the patient's current health condition, symptoms, and events leading up to their hospitalization.
6. **Physical Examination Findings**: Describes the healthcare provider's observations during the patient's physical examination, including vital signs, general appearance, and findings related to specific body systems (e.g., cardiovascular, respiratory).
7. **Diagnostic Test Results**: Includes laboratory test results (e.g., blood tests, imaging studies), electrocardiogram (ECG) findings, electroencephalogram (EEG) results, and other diagnostic tests performed during the patient's hospitalization.

8. **Hospital Course**: Describes the sequence of events and treatments administered during the patient's hospital stay, including procedures, medications, interventions, and any complications encountered.
9. **Medications:** Lists the medications prescribed to the patient on admission and discharge, including dosage, frequency, and route of administration.
10. **Discharge Diagnosis:** Specifies the primary diagnosis or diagnoses assigned to the patient upon discharge from the hospital.
11. **Discharge Condition**: Describes the patient's overall condition at the time of discharge, including their mental status, level of consciousness, activity status, and any ongoing care needs.
12. **Discharge Instructions**: Offers direction and suggestions for the patient's after-discharge care, including up-coming appointments, medication instructions, activity restrictions, and other relevant information.

    These features collectively provide a comprehensive overview of each patient's medical journey during their hospitalization, enabling various analyses, insights, and decision-making processes in clinical research and healthcare delivery.

```
Admission Date:  [**2115-2-22**]              Discharge Date:   [**2115-3-19**]

Date of Birth:  [**2078-8-9**]             Sex:   M

Service: MEDICINE

Allergies:
Vicodin

Attending:[**First Name3 (LF) 4891**]
Chief Complaint:
Post-cardiac arrest, asthma exacerbation

Major Surgical or Invasive Procedure:
Intubation
Removal of chest tubes placed at an outside hospital
R CVL placement


History of Present Illness:
Mr. [**Known lastname 3234**] is a 36 year old gentleman with a PMH signifciant
with dilated cardiomyopathy s/p AICD, asthma, and HTN admitted
to an OSH with dyspnea now admitted to the MICU after PEA arrest
x2. The patient initially presented to LGH ED with hypoxemic
respiratory distress. While at the OSH, he received CTX,
azithromycin, SC epinephrine, and solumedrol. While at the OSH,
he became confused and subsequently had an episode of PEA arrest
and was intubated. He received epinephrine, atropine, magnesium,
and bicarb. In addition, he had bilateral needle thoracostomies
with report of air return on the left, and he subsequently had
bilateral chest tubes placed.  After approximately 15-20 minutes
of rescucitation, he had ROSC. He received vecuronium and was
started on an epi gtt for asthma and a cooling protocol, and was
then transferred to [**Hospital1 18**] for further evaluation. Of note, the
patient was admitted to LGH in [**1-4**] for dyspnea, and was
subsequently diagnosed with a CAP and asthma treated with CTX
and azithromycin. Per his family, he has also had multiple
admissions this winter for asthma exacerbations.
```

Fig 1: Sample Dataset Record

4. Proposed Model Architecture

The architecture of Clinical BERT is build on the original BERT (Bidirectional Encoder Representations from Transformers) model [12], which includes multiple layers of Transformer blocks. However, Clinical BERT is specifically fine-tuned on large-scale clinical text data to capture domain-specific medical knowledge effectively.

4.1 Pre-training

Clinical BERT undergoes pre-training on a vast amount of clinical text, including electronic health records, medical literature, and other healthcare documents [13]. During pre-training, the model picks up on missing word predictions in sentences using bidirectional context and captures the contextual understanding of medical terminologies and their relationships.

4.2 Fine-tuning for Knowledge Discovery

After the pre-training phase, Clinical BERT is further fine-tuned for knowledge discovery tasks in healthcare. This fine-tuning process [14] involves training the model on specific datasets related to the target knowledge discovery and graph representation tasks. The fine-tuning process enables Clinical BERT to adapt its learned representations to the specific nuances and intricacies of the healthcare domain, allowing it to better capture and encode medical knowledge in a way that is tailored to the needs of knowledge discovery and graph representation.

4.3 Knowledge Graph Representation

One of the key strengths of leveraging Clinical BERT for knowledge discovery is its ability to construct comprehensive and meaningful knowledge graphs that represent interconnected medical concepts and relationships. By utilizing the rich contextual understanding encoded in Clinical BERT's representations, the generated knowledge graphs capture intricate dependencies and associations among clinical entities, providing a holistic view of medical knowledge.

The knowledge graphs derived from Clinical BERT can serve as valuable resources for various applications, including semantic search, clinical decision support systems, and predictive modelling. Furthermore, they facilitate efficient exploration and visualization of complex medical relationships, offering insights that can inform evidence-based healthcare practices.

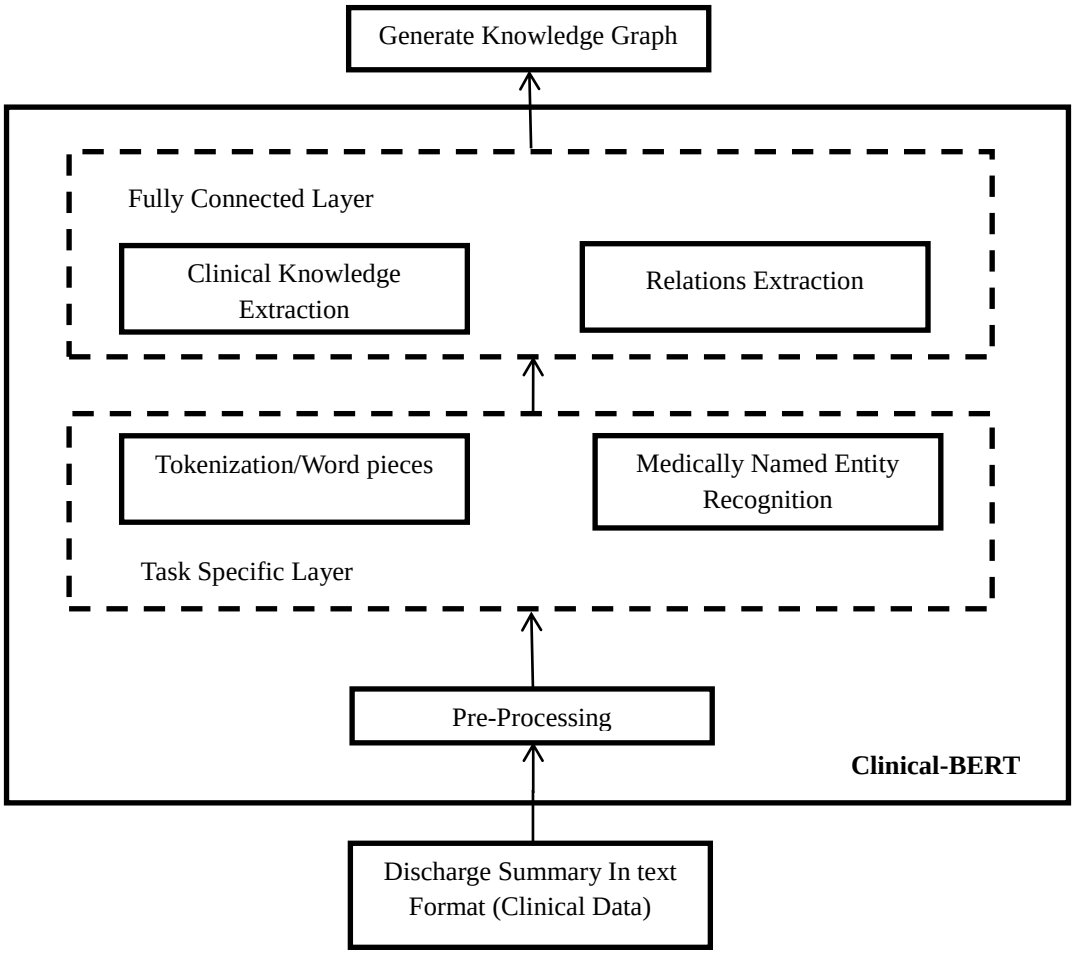

Fig 2. Architecture of our proposed model

In conclusion, the Fig2 shows detailed architecture of Clinical BERT, along with its fine-tuning for knowledge discovery and graph representation, equips it with the capabilities to revolutionize healthcare knowledge extraction and utilization. As the integration of advanced natural language processing models continues to evolve, the potential for Clinical BERT to enhance various facets of healthcare becomes increasingly apparent. It can significantly improve the accuracy of clinical decision support systems by providing precise and contextually relevant insights, thus aiding healthcare professionals in making better-informed decisions. Furthermore, the ability of Clinical BERT to construct detailed and comprehensive knowledge graphs allows for improved semantic search capabilities, enabling users to find relevant medical information quickly and efficiently. This enhances the accessibility of critical data, supporting both clinical and research efforts. In predictive modeling, the nuanced understanding of clinical concepts and relationships captured by Clinical BERT's representations can lead to more accurate predictions of patient outcomes, disease progression, and treatment responses. Overall, the advancements brought about by Clinical BERT in processing and understanding clinical text data hold the promise of significantly advancing the quality, efficiency, and efficacy of healthcare delivery and research.

5. Comparative Analysis

The F1-Score of the proposed clinical BERT model for acquiring clinical knowledge from text data has been evaluated against existing clinical models. The results, illustrated in Fig3 and detailed in Table 1, indicate that transformer models outperform CNN and LSTM based approaches in clinical tasks, as shown in the graph.

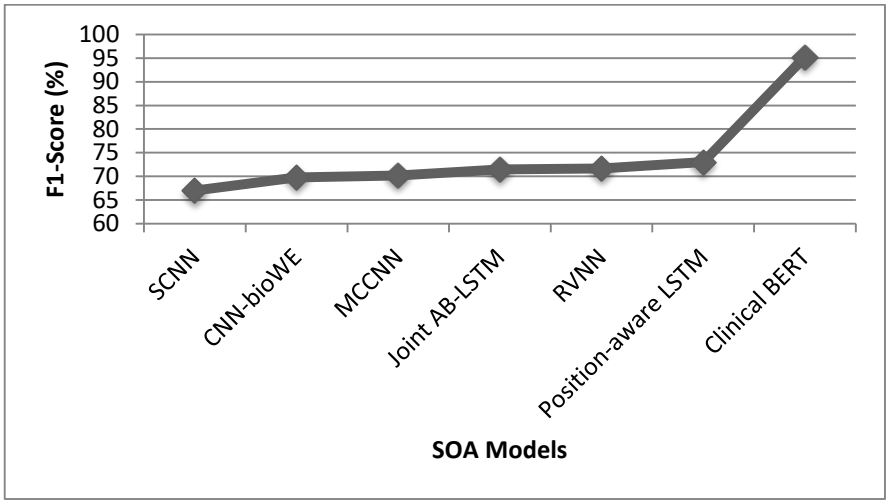

Fig 3. Comparative Analysis- Clinical-BERT model with SOA models

Table 1: Clinical-BERT's F1-Score comparison with SOA models

Sl.No.	Author Name & Reference No.	Existing Models	F1-Score(%)	
			Existing Models	Clinical-BERT
	Zhao Z et al. [15]	SCNN	67	95.2
	Liu S et al. [16]	CNN-bioWE	69.8	
	Quan C et al. [17]	MCCNN	70.2	
	Sahu S. K. and AnandA. [18]	Joint AB-LSTM	71.5	
	Lim S et al. [19]	RVNN	71.7	
	Zhou D et al. [20]	Position-aware LSTM	73	

6. Results

Utilizing the advanced capabilities of the Clinical BERT model, the study delves into the automated retrieval of vital data from clinical discharge records. When the text exceeds 512 words or tokens, BERT processes the data by breaking it into chunks. The model identifies key entities and their relationships, specified externally based on the requirements, offering valuable insights into patient diagnoses, treatments, and outcomes. Fig4 below shows the knowledge extracted from the clinical discharge text file.

The work extends beyond simple data extraction by building a knowledge graph to visualize and understand the complex relationships between the identified entities. This graph acts as a structured representation of concepts, allowing for a deeper comprehension of the underlying connections within the data. Fig5 illustrates the knowledge graph created for the given dataset.

Processing Chunk 1
Processing Chunk 2
Processing Chunk 3
Processing Chunk 4
Processing Chunk 5
Processing Chunk 6
Processing Chunk 7
Processing Chunk 8
Processing Chunk 9
Processing Chunk 10
Processing Chunk 11
Processing Chunk 12
Processing Chunk 13
Processing Chunk 14
Processing Chunk 15
Processing Chunk 16
Processing Chunk 17
Processing Chunk 18
Processing Chunk 19
Processing Chunk 20
Processing Chunk 21
Processing Chunk 22
Processing Chunk 23
Processing Chunk 24
Processing Chunk 25
Processing Chunk 26
Processing Chunk 27
Processing Chunk 28
Processing Chunk 29
Processing Chunk 30
Processing Chunk 31
Processing Chunk 32
Processing Chunk 33
Processing Chunk 34
Processing Chunk 35
Processing Chunk 36
Processing Chunk 37
Admission Date: [**2115-2-22**]
Discharge Date: [**2115-3-19**]

Date of Birth: [**2078-8-9**] Sex: M

Service: MEDICINE

Allergies:
Vicodin

Attending:[**First Name3 (LF) 4891**]
Chief Complaint:
Post-cardiac arrest, asthma exacerbation

Major Surgical or Invasive Procedure:
Intubation
Removal of chest tubes placed at an outside hospital
R CVL placement

History of Present Illness:
Mr. [**Known lastname 3234**] is a 36 year old gentleman with a PMH signifciant
with dilated c
Patient Demographics: [**2078-8-9**]
Chief Complaint: Post-cardiac arrest, asthma exacerbation
History of Present Illness: Mr. [**Known lastname 3234**] is a 36 year old gentleman with a PMH signi
fciant
with dilated c
Past Medical History: Asthma
Dilated cardiomyopathy
Multiple admissions

Fig 4. Knowledge Extracted

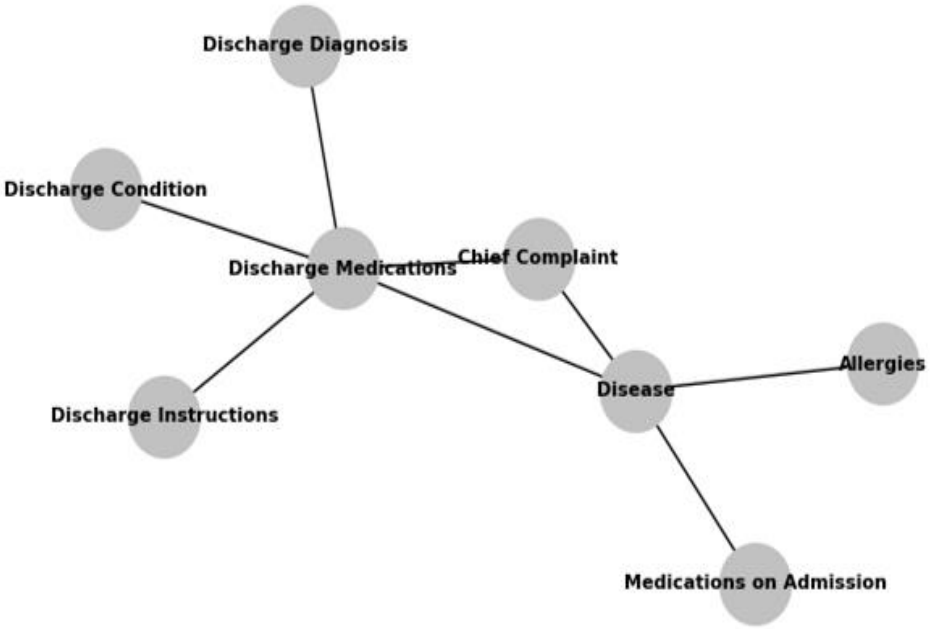

Fig 5. Knowledge Graph

Each node within the knowledge graph represents specific data extracted from the clinical discharge files. These nodes act as central information points, offering detailed insights into patient demographics, medical history, treatment procedures, and other critical aspects of the healthcare journey. Fig6(a) and Fig6(b) below illustrate the data encapsulated at each node of the graph generated in the previous step, as mentioned in Fig5..

```
Node: Disease
Details: Asthma
Dilated cardiomyopathy
Multiple admissions for dyspnea this winter ([**1-26**]).
Anxiety/depression
CKD
HLD
Obesity
HTN

Node: Allergies
Details: Vicodin

Node: Medications on Admission:
Details: Carvedilol 25 [**Hospital1 **]
Lasix 80 mg po bid
Xanax 0.25 mg 1-2 tabs prn
albuterol MDI
Ibuprofen prn
Benadryl prn
Advair diskus
Lsinopril 40 daily

Node: Chief Complaint
Details: Post-cardiac arrest, asthma exacerbation

Node: Discharge Medications
Details: 1. bisacodyl 5 mg Tablet, Delayed Release (E.C.) Sig: Two (2)
Tablet, Delayed Release (E.C.) PO DAILY (Daily) as needed for
Constipation.
2. senna 8.6 mg Tablet Sig: One (1) Tablet PO BID (2 times a
day) as needed for Constipation.
3. acetaminophen 325 mg Tablet Sig: Two (2) Tablet PO Q6H (every
6 hours) as needed for pain/fever.
4. carvedilol 12.5 mg Tablet Sig: Two (2) Tablet PO BID (2 times
a day).
5. docusate sodium 100 mg Capsule Sig: One (1) Capsule PO BID (2
times a day).
6. furosemide 40 mg Tablet Sig: Two (2) Tablet PO BID (2 times a
day).
7. lisinopril 10 mg Tablet Sig: Two (2) Tablet PO DAILY (Daily).
```

Fig 6(a). Data encapsulated at each respective nodes

8. olanzapine 5 mg Tablet, Rapid Dissolve Sig: [**11-25**] Tablet, Rapid Dissolves PO QHS (once a day (at bedtime)) as needed for sleep.

9. calcium carbonate 200 mg (500 mg) Tablet, Chewable Sig: One (1) Tablet, Chewable PO BID (2 times a day).
10. cholecalciferol (vitamin D3) 400 unit Tablet Sig: One (1) Tablet PO DAILY (Daily).
11. acetaminophen 500 mg Tablet Sig: Two (2) Tablet PO TID (3 times a day) as needed for pain/fever.
12. lidocaine 5 %(700 mg/patch) Adhesive Patch, Medicated Sig: One (1) Adhesive Patch, Medicated Topical DAILY (Daily): 12 hours on and 12 hours off every 24 hour period.
13. ipratropium bromide 0.02 % Solution Sig: One (1) neb Inhalation every six (6) hours.
14. albuterol sulfate 2.5 mg /3 mL (0.083 %) Solution for Nebulization Sig: One (1) neb Inhalation every six (6) hours.
15. albuterol sulfate 2.5 mg /3 mL (0.083 %) Solution for Nebulization Sig: One (1) neb Inhalation Q2H (every 2 hours) as needed for SOB.
16. topiramate 25 mg Tablet Sig: One (1) Tablet PO BID (2 times a day) for 3 days: 1 [**Hospital1 **] until [**3-22**] PM then increase to 2 tablets [**Hospital1 **] for 7 days then 3 tablets [**Hospital1 **] ongoing.
17. tramadol 50 mg Tablet Sig: One (1) Tablet PO Q6H (every 6 hours) as needed for back pain.
18. fluticasone-salmeterol 250-50 mcg/dose Disk with Device Sig: One (1) inh Inhalation [**Hospital1 **] (2 times a day).
19. lorazepam 2 mg/mL Syringe Sig: 1-2 mg Injection twice a day as needed for seizure that last longer than 5 minutes.

Node: **Discharge Diagnosis**
Details: Anoxic Brain Injury s/p PEA arrest x2
Status Asthmaticus
Ventilator Associated Pneumonia
Chronic Systolic Heart Failure
L1 compression fracture
Seizures after hypoxic brain injury

Node: **Discharge Condition**
Details: Mental Status: Confused - sometimes.
Level of Consciousness: Alert and interactive.
Activity Status: Ambulatory - requires assistance or aid (walker or cane) because he has poor motor planning

Node: **Discharge Instructions**
Details: You came to the hospital after having a cardiac arrest and an asthma exacerbation. You had another cardiac arrest in our hospital and were admitted to the MICU. You required intubation but were able to wean off the machine and breathe on your own.
We treated you for pneumonia and asthma. Your mental status slowly improved, though you did have 2 seizures, last on [**3-18**].
You were started ons eizure medications for this.
.
Please take your medications as prescribed and follow up with your doctors [**Name5 (PTitle) 7928**].

Node: **Medications on Admission**
Details: No details

Fig 6(b). Data encapsulated at each respective nodes

Complementing the knowledge graph insights, the study employs statistical analysis to quantify patterns and correlations in clinical data. Fig7 depicts the word frequency distribution from discharge summaries, highlighting common themes and identifying anomalies, such as significant terms or outliers. This analysis helps establish baseline expectations for word occurrences, aiding in textual data interpretation.

Fig8 illustrates tablet name frequencies, revealing "Pantoprazole 40 mg" and "Sodium 100 mg" as the most common, each appearing 69 and 68 times, respectively. Other frequently mentioned medications include "Aspirin 81 mg" and "Tartrate 25 mg," each with 52 occurrences. This analysis enhances evidence-based decision-making and patient care by identifying prevalent medications, usage trends, and prescribing patterns, contributing to more effective medication management and treatment protocols.

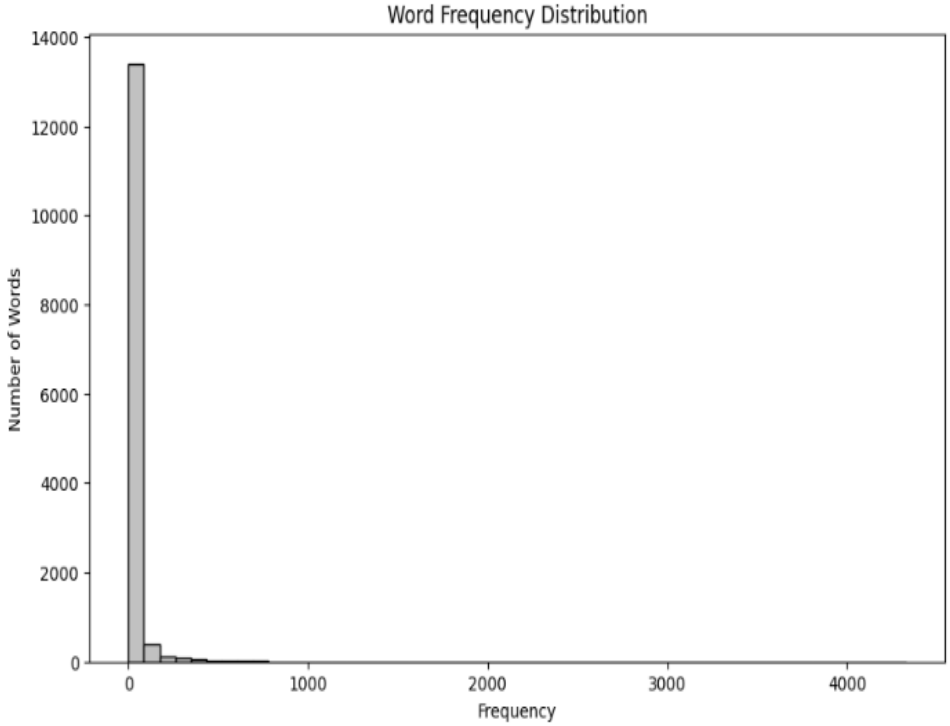

Fig 7. Clinical Word Frequency Distribution

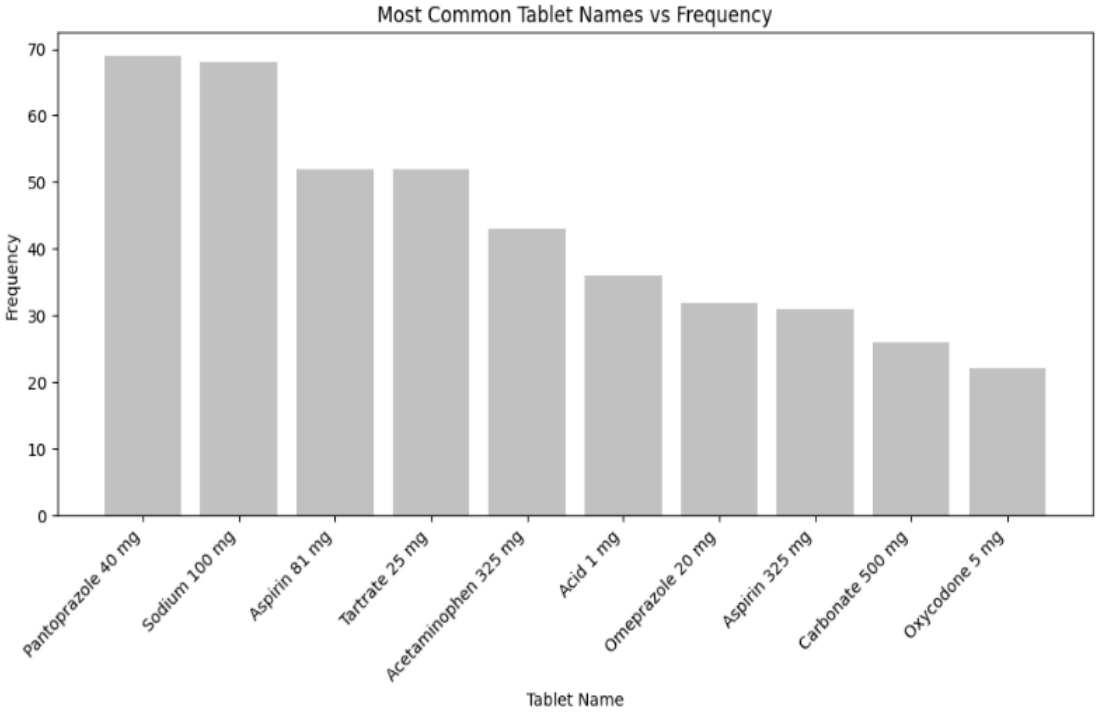

Fig 8. Most Common Tablet Names vs Frequency

7. Hyperparameter Tuning

Below Table 7.1 shows the different parameter configurations, corresponding accuracy, and duration time for training the BERT model. By fine-tuning these hyper-parameters, the Clinical BERT model can be optimized for better performance in understanding and processing clinical text data, ultimately improving its utility in healthcare applications.

Table 7.1: Hyperparameter Configurations

Hidden Size	Num Heads	Input Size	Batch Size	Num Epochs	Accuracy	Duration Time (minutes)
768	12	128	17	10	95.2%	120
512	8	128	32	15	96.7%	180
1024	16	256	32	10	97.8%	200
512	12	256	64	12	98.5%	240

8. Conclusion

In conclusion, our study underscores the transformative potential of leveraging Clinical BERT and graph-based representations for knowledge discovery in healthcare. Through our comprehensive approach, we have demonstrated the efficacy of employing state-of-the-art NLP techniques to extract valuable insights from clinical text data. By fine-tuning the Clinical BERT model and utilizing NetworkX for graph construction and analysis, we have facilitated the creation of a structured representation of medical knowledge in the form of a Knowledge Graph. This graph enables researchers, clinicians, and data scientists to navigate and explore complex relationships between medical concepts, eventually resulting in a greater comprehension of the mechanisms behind disease, treatment modalities, and patient outcomes.Our findings highlight the importance of interdisciplinary collaboration between NLP experts, healthcare professionals, and data scientists in harnessing the potential of advanced computational techniques to address critical challenges in healthcare. Moving forward, we plan to continue refinement and expanding of our approach to incorporate extra sources of clinical data, enhance the accuracy and granularity of entity extraction and relationship inference, and enable real-time decision support in clinical settings. Overall, our study underscores the transformative impact of integrating NLP and graph-based representations in healthcare, paving the way for more informed decision-making, improved patient care, and advancements in medical research. As we continue to push the boundaries of innovation in healthcare informatics, we are confident that our approach will contribute to the on-going efforts to enhance healthcare delivery and ultimately improve patient outcomes.

Data Availability Statement

The data that support the findings of this study were provided by Partners HealthCare System, Inc., through its i2b2 National Center for Biomedical Computing under a confidentiality agreement. Due to the nature of this agreement, the data cannot be shared publicly. Interested researchers may visit https://portal.dbmi.hms.harvard.edu to inquire about access to the data under similar terms and conditions.

References

1. Babu A, Boddu SB (2024) BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. Explor Res Clin Soc Pharm. 2024 Feb 15;13:100419. doi: 10.1016/j.rcsop.2024.100419. PMID: 38495953; PMCID: PMC10940906
2. Rasmy, L., Xiang, Y., Xie, Z. et al. (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. 4, 86. https://doi.org/10.1038/s41746-021-00455-y
3. Park Y-J, Lee M-a, Yang G-J, Park SJ, Sohn C-B (2023) Web Interface of NER and RE with BERT for Biomedical Text Mining. Applied Sciences. 13(8):5163. https://doi.org/10.3390/app13085163
4. Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., & El Asri, B. (2021). BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update, 1, 100042
5. Hassanpour, S., Langlotz, C. P., & Amrhein, T. J. (2016). Classification of radiology reports for falls in an emergency department electronic medical record: Concordance with experts. Journal of Biomedical Informatics, 64, 106-115. doi:10.1016/j.jbi.2016.09.005
6. Huang, Z., Xu, W., & Yu, K. (2019). Clinical entity recognition using hybrid approaches. BMC Medical Informatics and Decision Making, 19(Suppl 7), 232. doi:10.1186/s12911-019-0946-4
7. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 72-78. doi:10.18653/v1/w19-1909
8. Zhou, T., Wang, Z., Wang, L., Li, C., & Zeng, Z. (2020). Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. Journal of Biomedical Informatics, 111, 103569. doi:10.1016/j.jbi.2020.103569
9. Zhang, X., Xiao, C., Lu, H., Li, Y., Wang, F., & Wang, H. (2021). Building a knowledge graph of electronic health records for clinical decision support. Journal of Biomedical Informatics, 113, 103656. doi:10.1016/j.jbi.2020.103656
10. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö (2019) Cohort selection for clinical trials: n2c2 2018 shared task track 1. Journal of the American Medical Informatics Association. 26(11):1163–1171. https://doi.org/10.1093/jamia/ocz163
11. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner Ö (2019) 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association ;27(1):3-12. https://doi.org/10.1093/jamia/ocz166

12. Liu, X., Chen, H., & Zheng, J. (2019). Using sentiment analysis to explore online vaccination discussions and attitudes of Chinese parents. Human Vaccines & Immunotherapeutics, 15(9), 2099-2106. doi:10.1080/21645515.2019.1603656
13. Li, X., Tang, B., Jiang, M., Denny, J. C., Xu, H., & Jiang, Z. (2018). Developing a time-sensitive version of clinicalBERT for improving temporal relation extraction from clinical notes. Journal of the American Medical Informatics Association, 25(7), 963-969. doi:10.1093/jamia/ocy037
14. Zheng, J., Chaudhari, A. S., Rajan, S., Halabi, S. S., & von Deneen, K. M. (2020). Deep learning framework for multimodal disease progression prediction from Alzheimer's disease data. Journal of Medical Imaging, 7(1), 014502. doi:10.1117/1.JMI.7.1.014502
15. Zhao, Z., Yang, Z., Luo, L., Lin, H. & Wang, J (2016) Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics 32, 3444–3453
16. Liu, S., Tang, B., Chen, Q. & Wang, X (2016) Drug-drug interaction extraction via convolutional neural networks. Comput. Math. Methods Med. 2016, 6918381
17. Quan, C., Hua, L., Sun, X. & Bai, W (2016) Multichannel convolutional neural network for biological relation extraction. Biomed Res. Int. 2016, 1850404
18. Sahu, S. K. &Anand, A (2018) Drug–drug interaction extraction from biomedical texts using long short-term memory network. J. Biomed. Inform. 86, 15–24
19. Lim, S., Lee, K. & Kang, J (2018) Drug drug interaction extraction from the literature using a recursive neural network. PLoS ONE 13, e0190926
20. Zhou, D., Miao, L. & He, Y (2018) Position-aware deep multi-task learning for drug–drug interaction extraction. Artif. Intell. Med. 87, 1–8