KNOWLEDGE NET: AN AUTOMATED CLINICAL KNOWLEDGE GRAPH GENERATION FRAMEWORK FOR EVIDENCE BASED MEDICINE

by

FAKHARE ALAM

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE AND INFORMATICS

2023

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Khalid Mahmood Malik, Ph.D., Chair
Mohammad-Reza Siadat, Ph.D.
Tianle Ma, Ph.D.
Ramin Homayouni, Ph.D.
David Corliss, Ph.D.

To my beloved daughters

Maryam and Sara

To my wife Zakiya

who stood by me as a pillar of support through this journey

To my parents who are inspiration behind this work

ACKNOWLEDGMENTS

Success is never achieved alone, and I owe a debt of gratitude to those who have made it possible for me to succeed. I am deeply thankful to my family, who have been a constant source of inspiration and provided their unwavering support throughout my research journey at Oakland University. Their devotion and sacrifices have instilled a sense of dedication and provided the confidence to pursue my dreams, and I am grateful for their role in shaping my foundation during both challenging and successful times.

I would also like to express my heartfelt appreciation to my advisor, Dr. Khalid Mahmood Malik, who has been an exceptional guide and mentor throughout my research. His encouragement and guidance have pushed me to achieve my goals, and his knowledge and assistance have helped me progress positively. In addition, I am grateful to my doctoral advisory committee members, Dr. Mohammad-Reza Siadat, Dr. Tianle Ma, Dr. Ramin Homayouni, and Dr. David Corliss, for their invaluable support, insightful feedback, and helpful suggestions throughout the research process. Their guidance has been instrumental in shaping the direction of my research.

I am also deeply grateful for the excellent collaboration and teamwork of my co-authors, Dr. Muhammad Afzal, Dr. Maqbool Hussain, Dr. Madan Krishnamurthy, and Hamed Babaei. I extend my sincere gratitude to Dr. Ghaus Malik for his pioneering work in neuroscience, especially in the field of brain aneurysms, which has deeply influenced my research framework for evidence-based medicine. Furthermore, I am extremely thankful for the support and encouragement provided by General Motors data science manager, Aaron Wolf, throughout this process.

Lastly, I would like to thank the CSE faculty, graduate office, and OU Kresge Library Writing Center for their invaluable assistance and guidance during my time at OU.

Fakhare Alam

PREFACE

The pursuit of our dreams can lead us on a winding path through different domains and disciplines. I always had a dream of pursuing a research career in the interdisciplinary area of computer science, artificial intelligence, and other prominent domains such as geography, medicine, and agriculture. A previous attempt I made to do research in geoinformatics at one of the prestigious institutes in India got rescinded due to unavoidable circumstances and I had to leave the institute and join the information technology industry. However, the passion for inter-disciplinary AI research has always intrigued me and I was motivated to start again. In 2015, I moved to the United States and started working as a Data Scientist in the automotive industry. Using predictive and prescriptive analytics, I worked on solving issues related to warranty, plant absenteeism, and vehicle complexity reduction, which in turn led me to consider the possibility of applying my data science and analytical skills to the healthcare industry and solving problems that are affecting people more directly than they do cars. After reading numerous healthcare research articles and doing some exploratory research, I discovered that there is wealth of valuable information available in a vast corpus of medical research papers, but no efficient framework for automating the curation and organization of these knowledge sources exists. Hence the key motivation for my research was to build an automated framework to extract evidence-based information, enriching the extracted knowledge with hospital data and enable healthcare professionals in making data-driven clinical decisions. The structured and unstructured data provided by Dr. Ghaus Malik regarding cerebral aneurysms in patients proved to be a boon to me since I could test the

automated framework of curating knowledge by taking IAs as a case study. As COVID-19 affected a wide range of people's lives, it motivated me to curate information, extract knowledge, and test this framework on a large corpus of research articles related to COVID-19.

The current construction frameworks for Knowledge Graphs (KGs) fail to generate relevant information for evidence-based practitioners. The organization of subgraphs lacks specificity to topics and is not suitable for evidence-based PICO queries. Manual or semi-automated processes used to build these KGs make them incapable of adapting to new domains and incorporating changing information, resulting in loss of relevance and missing important evidence over time. Neglecting temporal information and the dynamic nature of entities and relations can lead to erroneous information extraction and suboptimal decision-making. The proposed automated clinical knowledge graph generation framework addresses these issues by extracting knowledge from a huge corpus of unstructured data, integrating structured patients' data, and semantically enriching the knowledge with specific and generic ontologies. In this framework, evolving knowledge can be automatically incorporated and processed to generate new evidence. The architecture of the framework is tested on two types of healthcare datasets, IAs and COVID and is flexible enough to be applied to other domains such as the food supply chain, dietary recommendations, and agriculture.

The research was partially funded by the Brain Aneurysm Foundation USA and the Deputyship for Research & Innovation, Ministry of Education, Saudi Arabia. The project received invaluable support from Henry Ford Health Systems, Bloomfield Hills, Michigan, which provided data and domain expert assistance to validate the results. Dr.

Ghaus Malik invested a considerable amount of resources in gathering and formatting the unstructured clinical notes, collected over several years, to create valuable structured data used in this research. I greatly appreciate his philanthropic insight and empathy towards the research world of aneurysms. I sincerely thank my advisor for his kind words, advice, and guidance throughout my research.

ABSTRACT

KNOWLEDGE NET: AN AUTOMATED CLINICAL KNOWLEDGE GRAPH GENERATION FRAMEWORK FOR EVIDENCE BASED MEDICINE

by

FAKHARE ALAM

Adviser: Khalid Mahmood Malik, Ph.D.

To practice the evidence-based medicine, clinicians are interested to find the most suitable research for the clinical decision making. The use of knowledge graphs (KGs) and Neuro-Symbolic methods to integrate and analyze complex and heterogeneous healthcare data is critical to enable evidence-based treatment in clinical decision support systems (CDSS). Healthcare generates a vast amount of data, including electronic health records (EHRs), medical images, genetic information, research papers, and clinical guidelines. Neuro-symbolic AI can leverage its neural network component to process unstructured data, while using symbolic reasoning to interpret the data and make logical inferences. It also enables a deeper understanding of patient data, leading to more accurate diagnoses, personalized treatment plans, and improved patient outcomes. By incorporating symbolic reasoning, Neuro-Symbolic AI systems can provide explanations for their outputs, making them more transparent and interpretable. To enable Neuro-Symbolic AI in healthcare, large-scale KGs play a pivotal role as it can integrate heterogeneous and big healthcare data including medical ontologies, clinical guidelines, drug databases, patient records, and research literature.

The existing KG construction frameworks are not fully automated and predominantly carried out using manual or semi-automated approach, requiring substantial effort and expertise. The challenges encompass identifying knowledge sources, disambiguating concepts in context, enriching semantics, determining relationships, and conducting inferential reasoning. Automating the extraction of coherent knowledge and constructing KGs from diverse data forms remains a longstanding goal in AI research. Also, the current frameworks for constructing KGs fail to generate KGs that provide relevant information for evidence-based practitioners. This is because the organization of constructed subgraphs is neither topic-specific nor evidence-based PICO (Participants/Problem P, Intervention-I, Comparison C, Outcome O) query-friendly. These KGs, built through manual or semi-automated processes, are incapable of adapting to new domains and incorporating the constantly changing information into their knowledge base. Consequently, they gradually lose relevance over time and miss out on important evidence. Thus, ignoring temporal information and failing to incorporate dynamic nature of entities and relations can lead to erroneous information extraction and suboptimal decision-making.

This dissertation proposes fully automated knowledge graph curation framework to curate information and create KG of different clinical domains by employing concept extraction, semantic enrichment, optimized clustering using Neuro-Symbolic approach, and state of art Recurrent Neural Networks (RNNs) with BioBERT based encoded representation to categorize PICO elements and predict relationships between concepts using huge corpus of publicly available literature on COVID-19 and cerebral aneurysms.

The evaluation shows that the proposed framework achieves significant improvement over baseline models and has 93 %, and 82 % accuracy on aneurysm and COVID data set respectively for PICO classification. The Neuro-Symbolic clustering approach outperforms traditional baseline models by 43 % and achieves average precision of 88 % across all identified clusters. Also, the relationship extraction module has an accuracy of 96 % with precision and recall being 92 %, and 90 % respectively.

The incorporation of domain-specific and language models has proven to enhance the performance of machine learning models, particularly in the context of Neuro-Symbolic clustering, PICO classification, and relation extraction. The integration of deep learning and symbolic reasoning techniques has demonstrated significant improvements in clustering performance, especially in biomedical research domains. The utilization of the BioBERT embedded layer and LSTM model has notably boosted the accuracy of PICO classification tasks by 11% for both the COVID-19 dataset and cerebral aneurysm dataset. Furthermore, when BioBERT is combined with Bi-LSTM and CNN, the performance of the RE model also experiences substantial enhancements. Future work will focus on parallelizing the data processing pipeline to enhance the efficiency and scalability of the knowledge graph framework, while also developing an interactive user interface for visualization. Additionally, efforts will be dedicated to extending the framework's application across diverse domains such as the food supply chain, dietary recommendations, agriculture, and fisheries, addressing unique challenges and expanding its impact. This expansion aims to advance multiple industries and leverage the potential benefits of the approach in various domains.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

KG	Knowledge Graph
PICO	Patient/Problem Intervention Comparison Outcome
BERT	Bidirectional Encoder Representations from Transformer
NLP	Natural Language Processing
IA	Intracranial Aneurysm
CB	Cerebral Aneurysm
SAH	Subarachnoid Hemorrhage
CDSS	Clinical Decision Support Systems
SPARQL	SPARQL Protocol and RDF Query Language
AI	Artificial Intelligence
GHKG	Google Healthcare Knowledge Graph
SNOMED	Systemized Nomenclature of Medicine
SAH	Subarachnoid Hemorrhage
ARDS	Acute Respiratory Distress Syndrome
DBSCAN	Density Based Spatial Clustering of Applicants with Noise
LDA	Latent Dirichlet Allocation
NER	Named Entity Recognition
CNN	Convolution Neural Network
RNN	Recurrent Neural Network
EHR	Electronic Health Record
CKG	COVID-19 Knowledge Graph

LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
LM	Language Model
SDS	Supervised Distance Supervision
CORD-19	COVID-19 Open Research Dataset
MIL	Multi-Instance Learning
UMLS	Unified Medical Language System
OBIE	Ontology Based Information Extraction
NLTK	Natural Language Tool Kit
URL	Uniform Resource Locator
PCA	Principal Component Analysis
MEDDRA	Medical Dictionary for Regulatory Activities
NBO	Neuro Behavior Ontology
NIFSTD	Neuroscience Information Framework Standard
LR	Logistics Regression
RF	Random Forest
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
KNN	K-Nearest Neighbors
LSA	Latent Semantic Analysis
TF-IDF	Term Frequency Inverse Document Frequency

SVD	Singular Value Decomposition
ACoA	Anterior Communication Artery
HTML	Hyper Text Markup Language

CHAPTER ONE

INTRODUCTION

The rapid growth of Knowledge Graph (KG) in recent years has been indicative of a resurgence in knowledge engineering. The combination of KGs and Neuro-Symbolic AI presents an immensely promising opportunity for pushing the boundaries of knowledge representation, reasoning, and inference. As new information emerges, Neuro-Symbolic AI techniques can adapt the KGs to incorporate the latest data, ensuring their relevance and timeliness. This dynamic knowledge representation enables more accurate information extraction and facilitates better decision-making in domains such as healthcare, where up-to-date information is critical. In many real-world applications, KGs have been found to be the primary source of information extraction, including text interpretation, recommendation systems, and answering natural language questions. Neuro-Symbolic AI techniques have seen significant application in various domains including healthcare, to provide reasoning and explanation [1,2]. In recent times, a number of knowledge bases have been created, covering common sense knowledge such as Cyc [3], conceptNet [4], lexical knowledge such as WordNet [5], encyclopedia knowledge such as DBpedia [6], YAGO [7], WikiData [8] CN-Dbpedia [9], probabilistic taxonomy knowledge like Probase [10] and geographical knowledge GeoNames [11]. Incorporating the above-mentioned open KGs and infusing it with external data sources using Neuro-Symbolic techniques has enabled a new form of knowledge storage that can capture semantically related large datasets. KGs that are generic in nature, as well as open world KGs, cannot address all the real-world issues.

The use of KGs on published literature to distill usable information that Neuro-Symbolic AI models and expert-based systems could use is one of the most promising approaches to the data consumption problem; and also, it provides more explanations for AI techniques such as machine learning and deep learning. While deep learning has shown tremendous promise in various fields, relying on it alone for healthcare can be insufficient as these models lack interpretability, making it challenging to understand the reasoning behind their predictions or decisions. In healthcare, interpretability and explainability are crucial for gaining trust from healthcare professionals and patients. To overcome these limitations and achieve more robust and reliable healthcare solutions, a holistic approach that combines deep learning with other techniques, such as symbolic reasoning, domain knowledge using human-in-loop validation, and contextual understanding, is necessary. This fusion of approaches can provide a more comprehensive and interpretable framework for addressing the complexities and nuances of healthcare.

At present, healthcare KGs are constructed in a static and manual or semi-automated fashion requiring significant effort and expertise, and their relevance may erode over time as the underlying contextual and temporal information keeps changing with the discovery of new facts, leading to new understanding of complex phenomena. Identification of knowledge sources, disambiguation, and recognition of concepts in context, semantic enrichment and relationship determination of concepts, clustering information to generate topics, and inferential reasoning are some of the challenges associated with the creation of a knowledge base. The automatic extraction of coherent knowledge and construction of KGs from the different forms of data is challenging

and a long-standing goal in artificial intelligence research. Data and entities in original data sources continuously change and evolve over time as new knowledge is added through articles, scientific manuals, and procedures such as recipes, how-to guides, etc. The various factors that define entities and relationships change over time, the ignorance of temporal information and the dynamic nature of the entities can lead to incorrect information extraction and poor decisions being made as a result of these factors. Therefore, automated construction of KGs must be more resilient and robust in order to accommodate continuous influxes of new knowledge, include temporal and contextual information, and make implicit information explicit.

To facilitate the integration of external knowledge, symbolic reasoning, and incorporation of human-in-the-loop feedback, it is crucial to develop an automated approach for constructing KGs. Automated construction methods offer streamlined processes, reduced manual effort, and accelerated KG creation. Published clinical knowledge is an enormous amount of data and the knowledge is constantly evolving, where physicians are constantly attempting to find new evidence of diagnosis, treatment, and eventually cure. KG presents a light way of organization of biomedical facts for different purposes such as question answering, search engines, recommender systems, or medical assistants and it powers AI-driven discovery related to different drugs and diseases.

The existing healthcare KGs have issues pertaining to retrieval of relevant information within acceptable response time [19], as the subgraphs are not optimized, and enabled to provide evidence-based information retrieval using PICO elements (Participants/Problem P, Intervention-I, Comparison C, Outcome O). The application

of evidence-based medicine is crucial for delivering high-quality care in healthcare such as nursing, family medicine, etc., since it allows clinicians, physicians, and patients to evaluate available research and evidence, and apply data-backed solutions to make quantitative and qualitative decisions based on scientific results. Additionally, this knowledge can be used to provide evidence-based treatment options, improve treatment effectiveness, and ultimately help find a way to mitigate the disease by providing healthcare providers with a better understanding of the disease and its symptoms. However, updating the evidence-based decision support system using continuous evolving unstructured voluminous literature is challenging. In evidence-based clinical decision support systems, KG can be used to store and process scientific evidence and clinical results, enabling a deeper understanding of symptoms, underlying causes, possible diagnoses, analyzing historical data about certain diseases, and assisting experts in dealing with complex cases. However, it can be an enormous task to consume this vast amount of published knowledge and present it in machine understandable format. Enriching the subgraph with specific problems, intervention, and outcomes is key to delivering relevant search results and it reduces the search time too. However, the PICO classification is challenging due to the unavailability of clean, curated, and preprocessed text corpus specific to the domain. Additionally, as knowledge evolves, the underlying topics, and the accuracy of entities and relationship among changes, making it difficult to keep KGs contextual and current. As a result of continuous evolution, it is also more difficult and challenging to maintain the continuity of existing relationships, construction of new relationships from new knowledge, and embed it into existing KGs because there must be a mechanism to

store temporal information and creation of new topics. The traditional clustering techniques such as K-means lack context and have limitation in providing explanation of topic identified out of clustering result. With the rise of embedding specific to biomedical domains such as BioBERT [20], there is a need to develop Neuro-Symbolic clustering framework using deep learning and symbolic reasoning having an ability to process big data, capture contextual information and infuse knowledge from domain specific external sources to enable reasoning in the cluster identification, creation and optimization. By harnessing the Neuro-Symbolic approach, the optimized clustering technique allows for the generation of PICO-specific subgraphs within each identified topic, facilitating evidence-based treatment.

There is a recent interest in developing KGs specific to a wide variety of specific domains such as education, science, automotive, and healthcare [12-16] and specific problems such as impact analysis [17], root cause analysis [18], and feature engineering. Domain and problem specific KGs are important because they include entities that are relevant to the domain and are semantically related to it. The construction of KG for rare diseases such as cerebral aneurysm or once in lifetime pandemic such as COVID-19 holds immense value in advancing the understanding, research, and clinical management of such conditions. Rare diseases often lack comprehensive and readily available information due to their low prevalence and limited research focus, while a pandemic involves vast amounts of diverse and rapidly evolving information from various sources such as research papers, clinical trials, public health reports, and social media. A KG can help in organizing and structuring this information, making it easily accessible for researchers, policymakers, and

healthcare professionals. The disease specific KGs organize existing knowledge, including genetic factors, risk factors, clinical manifestations, diagnostic methods, treatment options, and outcomes. This structured representation enables researchers, clinicians, and patients to access and explore relevant information, promoting collaboration, knowledge sharing, and evidence-based decision-making. By capturing the complexities of a rare disease within a Knowledge Graph (KG), research advancements can be fostered, personalized medicine approaches can be facilitated, and ultimately, patient care and outcomes for this rare disease can be improved.

This dissertation attempts to solve the above-mentioned challenges and presents an automated knowledge graph curation framework to address the challenges mentioned above. The framework utilizes Neuro-Symbolic clustering techniques to construct topic-specific knowledge graphs, enriched with PICO capabilities for evidence-based support. It effectively handles the dynamic nature of information, entities, and relationships, while seamlessly incorporating new data sources to maintain relevance. The implementation methodology focuses on fundamental knowledge graph qualities such as accuracy, consistency, completeness, timeliness, availability, and scalability [21]. Previous work by the authors [22,23] has demonstrated the utility of knowledge graphs in assisting clinicians and scientists. However, this dissertation goes further by presenting an automated framework specifically tailored to the healthcare domain, utilizing COVID-19 and cerebral aneurysms as illustrative examples. It addresses challenges related to concept identification and disambiguation through established ontologies and leverages the Neuro-Symbolic approach for optimized topic clustering. The framework incorporates

hierarchical relationships using specific ontologies and deep learning-based machine learning models, with a particular emphasis on contextual embedding facilitated by a domain-specific representation layer. It enriches entities through semantic techniques and ensures consistency and completeness through the integration of multiple ontologies. The framework also allows for easy updates, scalability, and adaptability to evolving healthcare domains, thus maintaining a comprehensive and current knowledge base

In summary, this dissertation introduces a comprehensive methodology that combines concepts, taxonomic relationships, and non-taxonomic relationships to construct an accurate, adaptable, and scalable framework for building a KG. The proposed framework demonstrates the capability to incorporate and assimilate continuously evolving clinical knowledge. By integrating diverse types of relationships, the framework achieves a holistic representation of domain-specific information, enabling the KG to capture the complex interplay between concepts. The flexibility and scalability of the framework empower healthcare professionals and researchers to effectively navigate and utilize the KG, facilitating informed decision-making and fostering the integration of emerging clinical insights.

CHAPTER TWO

BACKGROUND

The use of KG in evidence-based clinical decision support systems could provide an ability to store and process scientific evidence and clinical results, enable in-depth understanding of symptoms, related causes, potential diagnoses, enable historical data analysis about specific diseases, and assist experts when confronted with complex patient cases. KGs have been proven to be an effective utility in this arena and are rooted in many healthcare applications such as patient diagnosis [24], possible disease treatments [25,26] associate relationship between biomolecules and diseases [27], for better data representation and knowledge inference by mapping relationships between the enormous variety and structure of healthcare data. Also, structuring this large graph-based data should make the query processing easy.

The central objective of this research is to develop an automated framework that can extract knowledge from unstructured data and structure it using knowledge graph-based topic modeling, while incorporating PICO elements for evidence to tackle the challenges of evidence-based decision support systems. The framework is designed to improve the accuracy and efficiency of curation and organization of vast knowledge sources available in medical research papers. The major components of the proposed automated knowledge graph curation framework are illustrated in **Fig.1.**

In the subsequent subsection, this dissertation provides a brief overview of cerebral aneurysm and COVID-19, two case studies used in this research.

Additionally, and the PICO framework, which is widely used to formulate clinical research questions, is discussed.

Furthermore, the background of the curation framework components such as clustering, topic modeling, knowledge graphs, relation extraction, and embedded models BERT (Bidirectional Encoder Representations from Transformers) and BioBERT (BERT for Biomedical Text Mining) are also discussed. These components play a crucial role in the automated curation process and help improve the accuracy of the results.

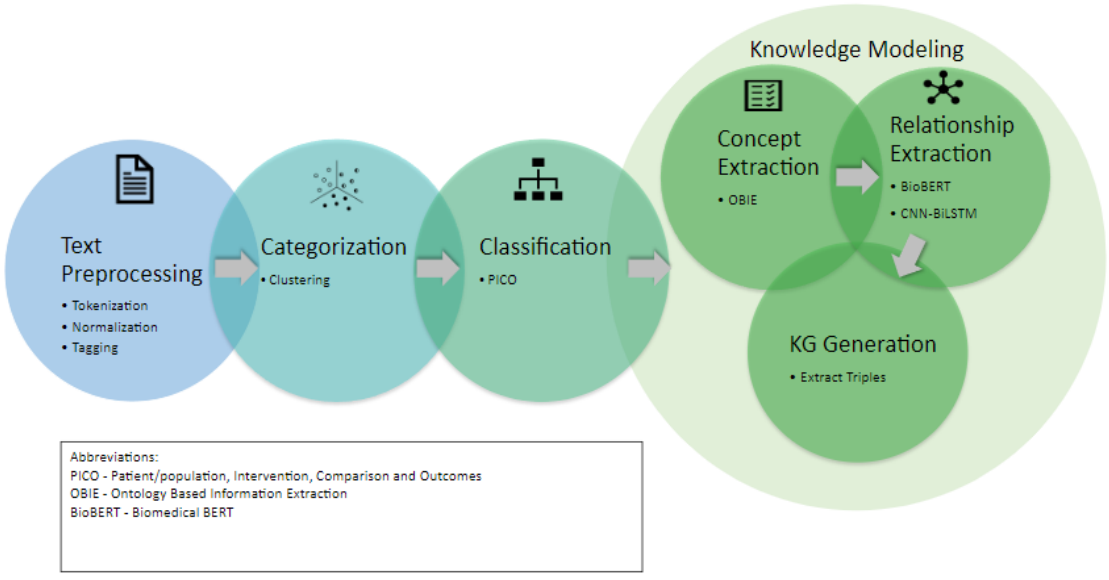

Fig. 1: Conceptual View of proposed Automated Knowledge Graph Curation Framework

2.1 Cerebral Aneurysm

The cerebral aneurysm, also called an intracranial aneurysm (IA) or brain aneurysm, occurs when a weak area in a blood vessel wall bulges or balloons out in the brain. When the vessel bulges, it can rupture, causing bleeding in the brain, which can cause a stroke or other serious complications. Brain aneurysms can occur anywhere in the

brain, and they often have no symptoms until they rupture, which can cause sudden and severe headaches, vision problems, and neurological deficits. Treatment options for brain aneurysms include surgery to clip or remove the aneurysm, or endovascular procedures to fill the aneurysm with tiny metal coils or stents to prevent rupture. The rupture of an aneurysm can be caused by many different factors, but the exact parameters that cause the rupture are not always known. However, there are several risk factors that are known to influence aneurysm. The risks of IA onset and its rupture are not well known are well known. Some of the known risk factors include smoking history, hypertension, family history, old age, gender, ethnicity, congenital abnormalities in the arteries, excessive alcohol and drug use, trauma, and other identified comorbid conditions including polycystic kidney disease.

2.2 COVID -19

COVID-19 is a highly contagious viral disease caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). It was first identified in Wuhan, China, in December 2019 and has since spread rapidly across the globe, leading to a global pandemic.

The primary mode of transmission of COVID-19 is through respiratory droplets when an infected person coughs, sneezes, or talks. The virus can also spread by touching contaminated surfaces and then touching the mouth, nose, or eyes.

The symptoms of COVID-19 range from mild to severe, including fever, cough, shortness of breath, fatigue, aches in the body, headache, loss of taste, sore throat, congestion, and diarrhea. In severe cases, the virus can lead to pneumonia, acute respiratory distress syndrome (ARDS), organ failure, and even death.

COVID-19 pandemic has had a profound impact on the world, leading to widespread illness, death, and economic disruption. The pandemic has also highlighted the importance of public health infrastructure, emergency preparedness, and scientific research in combating infectious diseases.

2.3 PICO Model

The PICO model is a commonly used framework in biomedical documents for structuring clinical questions and study design. The PICO framework allows the searcher to use specific terms to find clinically relevant evidence by organizing a clinical question. Without a well-focused question, it can be very difficult and time consuming to identify appropriate resources and search for relevant evidence. **PICO** stands for:

- **P**atient Problem, (or Population): The Patient element of the PICO model refers to the population of interest in the clinical question or study. This could include characteristics such as age, gender, medical history, and diagnosis.
- **I**ntervention: The Intervention element refers to the treatment or intervention being considered. This could be a drug, a medical procedure, or a behavioral intervention.
- **C**omparison or Control: The Comparison element is used to compare the intervention to an alternative or control group. This could be a placebo, an existing treatment, or a different dose or duration of the same intervention.
- **O**utcome: The Outcome element refers to the measure of interest or the goal of the study. This could include clinical outcomes such as survival or disease progression, as well as quality of life, patient satisfaction, or cost-effectiveness.

Table 1 illustrates the recommended structure for organizing information in the PICO classification system with an example.

Table 1 PICO formulation and Example [28]

PICO category	Question Formulation	Example
Patient OR **P**roblem OR **P**opulation	What are the patient's demographics such as age, gender and ethnicity? Or what is the or problem type?	Work-related neck muscle pain
Intervention	What type of intervention is being considered? For example, is this a medication of some type, or exercise, or rest?	Strength training of the painful muscle
Comparison or Control	Is there a comparison treatment to be considered? The comparison may be with another medication, another form of treatment such as exercise, or no treatment at all.	Rest
Outcome	What would be the desired effect you would like to see? What effects are not wanted? Are there any side effects involved with this form of testing or treatment?	Pain relief

Using the PICO model can help to clarify the key elements of a clinical question or study design, which in turn can help to identify the most relevant evidence and studies in a literature search. The PICO model is often used in evidence-based medicine and systematic reviews, where the goal is to identify and synthesize the best available evidence to answer a clinical question.

2.4 Documents Clustering

Document clustering is the process of grouping documents based on similarity. This is an important task in natural language processing and information retrieval, as it can help to organize large collections of documents and make them easier to navigate and analyze. Document clustering algorithms use a variety of techniques to measure similarity between documents, such as cosine similarity, Jaccard similarity, and

Euclidean distance. Once documents have been clustered, it is possible to perform various tasks on the resulting groups, such as summarization, classification, and topic modeling. Document clustering has many real-world applications, including in search engines, recommendation systems, and content management systems. There are several algorithms available for document clustering, such as K-means clustering, hierarchical clustering, and DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering [29].

The K-means algorithm is the most widely used clustering algorithm in many use cases. It is an unsupervised machine learning algorithm that partitions a set of documents into k clusters based on their similarity. K-means clustering works by randomly selecting k centroids and assigning each document to the nearest centroid. The centroid is then updated to the mean of all documents assigned to it. The process is repeated until convergence. K-means clustering is widely used for document clustering because it is fast and efficient.

2.5 Neuro -Symbolic approach for document clustering

Neuro Symbolic Clustering is an innovative approach that combines principles from both neurocomputing and symbolic reasoning to perform clustering tasks in a hybrid manner. It aims to integrate the strengths of neural networks, which excel at learning complex patterns from large datasets, with symbolic reasoning, which focuses on the interpretability and logical manipulation of knowledge. In neuro symbolic clustering, neural networks are utilized to learn the underlying patterns and representations of the data, capturing intricate relationships and dependencies. The network’s hidden layers act as feature extractors, transforming the input data into a latent

space representation. This learned representation is then combined with symbolic reasoning techniques, such as logical rules or knowledge graphs, to facilitate the clustering process.

The integration of symbolic reasoning enables neuro symbolic clustering to incorporate prior knowledge or domain-specific constraints into the clustering process. By incorporating symbolic rules or constraints, the clustering algorithm can enforce logical consistency and interpretability in the clustering results, leading to more meaningful and understandable cluster assignments.

The neuro symbolic clustering approach has gained attention in various domains, including machine learning, data mining, and artificial intelligence. It provides a promising avenue for tackling complex clustering problems where both accuracy and interpretability are essential. By fusing the power of neural networks with the expressiveness of symbolic reasoning, neuro symbolic clustering offers a unique solution to uncovering hidden patterns and structures in data while maintaining human-understandable results.

2.6 Topic Modeling

Topic modeling is a technique used in natural language processing to automatically discover latent topics that underlie a collection of documents. It is a form of unsupervised machine learning that can help to identify the main themes and patterns present in a corpus of text. One of the key advantages of topic modeling is that it can help to make large collections of text more manageable and easier to explore. For example, by using topic modeling to identify the main themes in a collection of news articles, it becomes possible to summarize the key trends and topics covered in the news.

The most used algorithm for topic modeling is Latent Dirichlet Allocation (LDA) [30]. LDA models each document as a distribution of topics, and each topic as a distribution of words. It works by iteratively estimating the distributions of topics and words until convergence. LDA can be used to uncover hidden themes and topics in a wide range of text data, from news articles to social media posts.

In Natural Language Processing (NLP), topic modeling is used for a wide range of applications, including sentiment analysis, text classification, and information retrieval. Topic modeling, for example, can be used to automatically classify documents into different categories or to identify the most relevant documents. It can also be applied to content recommendation systems. An analysis of a user's previous interactions with a system can guide the recommendation of new content that they are likely to find interesting.

2.7 Knowledge Graph

A knowledge graph is a way of representing knowledge in a structured form, using nodes and edges to represent entities and their relationships. It is a large semantic network, defined as a directed labeled graph $G\ (V, E)$ with a set of nodes V: $\{v_1, v_2 \ldots v_n\}$ and edges E: $\{e_1, e_2 \ldots e_n\}$. A node is a real-world entity, such as an object, an event, a situation, a place, or a person, while a boundary defines a relationship among them. Specifically in the healthcare, nodes can be diseases, drugs, or symptoms and edges could be defined a '*cures', 'treats', 'symptoms* . KGs organize raw information in a structured form, integrate knowledge across different sources by defining descriptions of entities and relationships that are both understandable by humans and machines. It stores rational facts and information in a centralized fashion, bridges data silos, and generates canonical

business knowledge models that can incorporate all data formats such as structured, semi-structured, and unstructured.

In biomedical domain, KGs are used to represent complex relationships between biological entities such as genes, proteins, and diseases. These relationships can be represented as edges between nodes, allowing researchers to navigate the complex web of relationships between biological entities in a way that is much easier than using traditional search methods. For example, a KG might represent the interactions between proteins involved in a particular biological pathway, or the relationships between genes that are associated with a particular disease. By organizing this information in a structured way, researchers can gain insights into the underlying mechanisms of biological processes and diseases.

One of the key advantages of knowledge graphs is their ability to integrate information from multiple sources. This is particularly important in biomedical domain, where data is often scattered across multiple databases and sources. By integrating this data into a KG, researchers can gain a more complete understanding of the relationships between different biological entities. Another advantage of knowledge graphs is their ability to support sophisticated querying and analysis. Researchers can use query languages such as SPARQL (SPARQL Protocol and RDF Query Language) to search for specific patterns in the knowledge graph, enabling them to uncover relationships and insights that would be difficult to find using traditional search methods.

By integrating information from multiple sources and supporting sophisticated querying and analysis, KGs possess the potential to revolutionize the understanding and exploration of vast amounts of data.

2.8 Relation Extraction

Relation Extraction (RE) is an NLP task that involves identifying and extracting semantic relationships between entities mentioned in text. These relationships can take various forms, such as cause-effect, part-whole, and entity-attribute relationships. In the biomedical domain, relation extraction has become increasingly important for mining valuable knowledge from vast amounts of scientific literature. It is used to identify valuable knowledge from vast amounts of scientific literature, such as drug-disease relationships, PPIs (protein-protein interactions), and gene-disease relationships. In the drug-disease relationships it helps analyzing large volumes of scientific literature and identify potential drug candidates for specific diseases or gain insights into the mechanism of action of existing drugs. For example, by using relation extraction techniques, researchers have identified potential new uses for existing drugs, such as repurposing antipsychotic drugs for the treatment of cancer. In the PPIs play a crucial role in many biological processes, and their identification can provide insights into the underlying mechanisms of diseases. By using relation extraction techniques, researchers can identify potential new PPIs and better understand the complex network of interactions between proteins in the human body.

RE is also used to identify other types of biomedical relationships, such as gene-disease relationships, protein-chemical interactions, and drug-target relationships. These relationships can be used to develop new treatments and therapies for diseases, and to gain a better understanding of the underlying biology of disease processes and develop new treatments and therapies to improve human health.

2.9 BERT and BioBERT

BERT is a natural language processing (NLP) model that was introduced by Google in 2018 [31]. It has since become one of the most widely used models in NLP, due to its ability to understand the context of words in a sentence.

BioBERT [20] is a variant of BERT that has been specifically trained on biomedical text. It was created by adapting the pre-training process of BERT to biomedical literature, using a large corpus of PubMed abstracts and full-text articles. BioBERT has been shown to outperform other biomedical NLP models on a range of tasks, including named entity recognition, relation extraction, and question answering.

In the medical domain, BioBERT has been used in a range of applications, including drug discovery, clinical decision support, and biomedical literature mining. For example, BioBERT has been used to identify drug-disease relationships by analyzing large volumes of scientific literature, enabling researchers to identify potential new drug candidates for specific diseases. It has also been used to develop clinical decision support systems that can analyze patient data and recommend treatments based on the latest medical research. By leveraging the power of BioBERT, these systems can provide more accurate and personalized recommendations, improving patient outcomes. In addition to drug discovery and clinical decision support, BioBERT has been used to analyze biomedical literature and identify new insights into the underlying biology of diseases. By using BioBERT to extract relationships between genes, proteins, and diseases, researchers can gain a better understanding of the complex network of interactions that drive disease processes.

CHAPTER THREE

RELATED WORK

This chapter provides a comprehensive overview of the literature on the utilization of KGs in the healthcare industry. It delves into the construction mechanism of KGs and explores different clustering mechanisms. Additionally, it investigates the integration of deep learning with external knowledge infusion through the Neuro-Symbolic approach.

The chapter also focuses on the methodology employed for relationship extraction, which entails identifying semantic relationships between entities. It sheds light on the challenges associated with relationship extraction, including the handling of ambiguous language, and examines various approaches used to overcome these challenges. Furthermore, the extraction of PICO elements, which play a crucial role in evidence-based decision support systems within the healthcare domain, is thoroughly examined. The chapter emphasizes the significance of PICO elements and explores different techniques for their extraction, including rule-based methods, machine learning, and deep learning techniques.

3.1 Knowledge Graph in Healthcare

In numerous studies, KGs have been constructed and used for clinical research and healthcare. KG has a great deal of potential for diagnosing diseases [32], especially in terms of identifying disease-symptom relationships [33]. Healthcare specific knowledge bases have gained traction as a result of the COVID-19 pandemic, and efforts

are being made to establish, enhance, and create disease specific knowledge bases since they provide a means of mining, storing, and analyzing vast amounts of multimodal, heterogeneous data. **Table 2** shows the review of recent healthcare knowledge generation by analyzing them on methodology such as reasoning, ML (Machine Learning) based, data sources and its limitations in terms of generalizability, PICO enabled, and topic specificity.

3.2 Knowledge Graph Construction Methodology and Data Sources

In traditional KG construction, Named Entity Recognition (NER) [42], entities disambiguation, and RE methods are involved, and it also depends on the underlying schema. Schema options include ontology schemas [7,43], schema free approaches, or hybrid approaches that combine schema-free and schema-based approaches. In recent times, KGs have been constructed using big data mining and NLP techniques, such as association mining, deep learning methods such as Convolution Neural Network (CNN), Recurrent Neural Network (RNN), machine learning models such as Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), XGBoost, Adaptive Boosting, and extract knowledge from different sources, including silos of data and entity repositories [44-46]. However, these methods need domain expert to validate the mined rules. There are also some built-in models available such as IBM Watsons [47], which uses KGs to enhance its cognitive computing capabilities by organizing and connecting vast amount of data, enabling it to understand, reason, and generate valuable insights but it relies heavily on structured data for optimal performance. AllenNLP [48], an open-source NLP library primarily focuses on deep learning techniques.

Table 2: Healthcare Knowledge Graph Review Summary- Implementation Methodology and Data Sources

Reference	Description	Data Sources	Characteristics
[34]	Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings	DeepDDI and Decagon	. Semi-Automated . Not PICO enabled . Less generalizable
[27]	multi-typed drug interaction prediction via efficient knowledge graph summarization	DrugBank Dataset	. Semi-Automated . Not PICO enabled . Less generalizable
[24]	Breast Cancer diagnosis from clinical ultrasound reports	Collected by clinicians Cancer Center of SUn Yat-sen University	. Automated manual feature creation . Not PICO enabled . Less generalizable
[35]	Investigating plausible reasoning over knowledge graphs for semantic based health data analytics	Medline	. Automated . Not PICO enabled . Less generalizable
[36]	Identification and classification diseases and symptoms across the collected medical records using classifiers such as Logistic Regression (LR), Naive Bayes (NB), Bayesian Inferences	Custom Data, Google Healthcare Knowledge Graph (GHKG)	. Automated human in loop . Not PICO enabled . Less generalizable
[37]	Multiple criteria decision-making approaches for healthcare management applications	PubMed, DrugBank, DrugBook, and UMLS	. Semi-Automated . Not PICO enabled . Less generalizable
[38]	Knowledge graph for traditional Chinese medicine health preservation	TCM data	. Semi-Automated . Not PICO enabled . Less generalizable
[39]	Italian healthcare knowledge graph	EHRs	. Manual . Not PICO enabled . Less generalizable
[40]	Personalized health knowledge graph	EHRs, weather data,	. Manual . Not PICO enabled . Generalizable
[41]	Real- world data medical knowledge graph	EMRs	. Manual . Not PICO enabled . Generalizable

Though it does not have a built-in knowledge graph like some other systems, AllenNLP provides functionalities and tools that allow users to leverage external knowledge graphs in their NLP workflows, but it has limited pretrained models. Another open-source library CoreNLP [49], offers a wide range of NLP functionalities such as tokenization, part-of-speech tagging, named entity recognition, dependency parsing, sentiment analysis, and coreference resolution. It provides comprehensive tools for NLP tasks, although it does not possess a built-in knowledge graph.

In healthcare, construction of domain specific KGs using the above-mentioned approach is also on the rise as it offers an effective way to extract meaningful information from different modalities and variety of data. Gatta, et al. [50] proposed an approach to mine medical data and store it in directed graphs for easy visualization and interpretation. Authors in [35] integrated ontology-based knowledge with reasoning for completeness by integrating various knowledge bases such as DrugBank, Disease Ontology, and semantic databases such as MEDLINE. Rastogi & Zaki [51] constructed the personalized KGs by using contextual information and combining it with knowledge bases. There has been effort made on constructing disease specific KGs to extract health-related symptoms and risk factors. Rotmensch, et.al [36] constructed KG by extracting disease and symptoms from Electronic Health Records (EHRs) and combining it with Google Health Knowledge Graph (GHKG). Likewise, Huang et.al [52] constructed KGs for depressions by utilizing PubMed, DrugBank and United Medical Language System (UMLS) library. Since the entities and relationship between them are changing, Ma et.al [53] attempted to construct a temporal KG for cognitive episodic memory by generalizing four static KGs vectors to 4-dimensional temporal KGs. Zhang, et al., [54] present a platform named

Health Knowledge Graph Builder that could be used to build disease specific KGs considering multiple sources of data including those of clinician feedback. The framework utilizes data from medical records to inspect new concepts/relationships and allows clinicians to annotate unstructured data based on existing clinical KGs. While this approach provides automated and manual knowledge representation and allows to add new diseases to existing KGs, the dataset used was limited in quantity. Also, the dependency on several sources and bias caused due to manual intervention might reduce the extensive usability of this framework. The work in [55] provides a case study on how to evaluate the accuracy of health KG models for disease-symptom relationships. Unstructured datasets from de-identified patient EHRs (273,174 EHRs) are extracted and each chronological patient file is identified as an object entity, GHKG is used as the comparison entity to evaluate disease-symptom edge on the KG. While the approach aims at availing accurate information on healthcare KGs, their study is limited in comparison with GHKG and unclear on the inferencing of outliers. Likewise, Li, et al.,[41] presents a series of methods to create a medical KG using a quadruplet structure with data extracted from large EHRs, laboratory reports, and radiology reports. NER is used to identify 9 different objects pertaining to diseases, symptoms, medicine, gender, age, exam, lab exam, lab item and surgery. Relationship is extracted by building triplets of two entities with a relation, and every relation is associated with 4 properties providing the quadruplet structure. The work provides a novel approach of quadruplet medical KG however, the KG build process is variable intensive and manual. Wise, et al., [56] present a COVID-19 Knowledge Graph (CKG) for retrieval of semantically similar articles. The proposed CKG uses a hybrid approach by combining semantic information with the

topological document information. Chen, et.al., [57] utilizes the rich CORD-19 (COVID-19 Open Research Dataset) dataset and maps the insights obtained from PubMed dataset using machine learning to present a KG that helps identify COVID-19 experts and concepts from published literature. Wang, et al.,[58] proposed COVID-KG, a knowledge discovery framework that uses semantic representation and ontologies to provide multimedia analysis of text and image data from literatures to aid in clinical Q&As and generate reports associated with COVID-19 drug administration based on a case study.

3.3 Knowledge Graph Relationship Extraction Methodology

Most of the work in relation extraction employed traditional machine learning and deep learning approaches in supervised and distant supervision fashion [59]. However, recently the pre-trained language models (PrLMs) [60] achieved state-of-the-art performances in many tasks of NLP including domain-specific REs. Pawar, et.al., [59] concluded that semi-supervised approaches are well suited for open domain RE systems due to their ease of scaling and extensibility in finding new relations.

The earlier approaches heavily relied on hand-crafted feature engineering. These methods could be categorized into lexical [61-64] and word embedding [65-67] features. Buscaldi, et.al.,[68] reported the popularity of SVM in machine learning models at the SemEval-2018 relationship extraction task. The machine learning based RE methods mostly suffers from feature quality and is unable to generalize well on many domains because of feature distinctions in different domains.

The latest studies in deep learning focus on extraction of deep relational features using CNN and RNN-based approaches, although attention mechanisms have been combined with CNN or RNN methods in a few cases. CNN-based methods as

implemented by Liu, et.al.,[69]; Zeng, et.al.,[70]; Lin, et.al.,[71] extract global features of relations with a sentence as high-level features for classification. RNN based methods as implemented by authors in [72-75] use forward and backward propagations to obtain sequential features by considering past and future sequences. Wang, et.al., [76] proposed CNN architecture, which relies on two levels of attention to better discern patterns in heterogeneous contexts. In addition, Lee, et.al.,[77] proposed a new end-to-end RNN model that incorporates entities and their latent types as features for RE. Similarly, Xiao & Liu, [78] introduced a hierarchical RNN using BiLSTM (Bidirectional Long Short-Term Memory), where it is capable of extracting information from raw sentences for RE. Attention based mechanisms allow more accurate feature extraction from concepts and can predict relations by analyzing the whole sentence, however it suffers from an out-of-vocabulary and local contextual representation issues since it is trained in a specific training set.

Recently researchers are also using transfer learning model such as PrLMs (Pre-trained Language Models) for relationship extraction. The basic architecture of PrLMs includes a stack of encoders, fully connected to a stack of decoders. Each encoder consists of a self-attentive component and anticipatory network blocks. Each decoder consists of a self-attention component, an encoder-decoder attention component, and an anticipation component block [79]. Domain-specific LMs (Language Models), such as BERT, Transformer-XL, and OpenAI's GPT-2 have been used for RE tasks in multiple domains. For example, Wu & He [80] proposed a model that exploits both the pre-trained BERT LM and information from the target entities for the relationship classification task. They fine-tuned BERT for the RE task and achieved higher results than CNN or RNN-

based approaches. The single BERT-based model in joint entity-relation extraction has been investigated by Eberts & Ulges [81] where they presented a BERT-based RE framework with a softmax classifier and showed that joint models with PrLMs are performing well. Harnoune, et al., [79] concluded that the BERT variant such as BioBERT allows faster learning process by using RNN and CNN based architecture in medical RE tasks. However, there are still areas for improvement. In another study, Kim, et.al.,[82] considered the BERT model for unsupervised relation extraction and introduced Co-BERT, an open information extraction service that uses the power of BERT models with minor adjustments in self attention. Authors conclude that the attention mechanism in fine-tuning LMs for domain-specific tasks plays a critical role, and this is the reason why BERT model representations perform well for domain specific relationship prediction.

3.4 Clustering using Neuro-Symbolic approach

Deep clustering offers several advantages over traditional methods. It excels in categorizing unstructured, high-dimensional data by effectively identifying intricate patterns, implementing dimensionality reduction, and accomplishing cluster assignments simultaneously. Consequently, it provides a highly scalable end-to-end framework. Nonetheless, certain challenges must be thoroughly evaluated, such as the quality of clustering based on document content and meaning, as well as the categorization of documents with multiple topics [83].

Various language modeling techniques, including vector space modeling, dimensionality reduction, topic modeling, and continuous space modeling, employ feature-based representations of documents in a collection. Further enhancing deep

clustering, segmenting documents into sections with distinct topics can be advantageous. For instance, in [84], authors leverage constrained clustering through seed words, while our current study employs external knowledge infusion utilizing a neural network facilitated by expert-labeled training data.

Deep clustering models have been widely sought out and implemented as significant tools for document analysis in extensive text corpora [83]. Authors in the work [85], present the application of deep learning clustering algorithms for unsupervised research in bioimaging, cancer genomics, and biomedical text mining. They utilize a variant of the recurrent neural network algorithm LSTM (long short-term memory networks) to achieve document clustering. Additionally, Davagdorj et.al., in [86] propose a biomedical document clustering framework based on BioBERT, a pre-trained language representation technique. Both methods possess their respective advantages in improving deep clustering analysis. LSTM excels in handling long-term dependencies and modeling complex sequential data, while BioBERT captures multi-directional context with faster training time [87]. In the current approach, the advantages of both methods are leveraged to implement a deep clustering framework that incorporates LSTM and BioBERT to achieve enhanced clustering accuracy.

Existing deep learning methods can benefit from the infusion of domain and conceptual knowledge [88]. Knowledge infusion involves the co-utilization of symbolic AI with data-driven AI, resulting in a class of neuro-symbolic AI methods called knowledge-infused learning (KiL) [89]. Neuro-symbolic approaches enhance the efficiency of clustering frameworks by combining symbolic reasoning with neural perception. Work by Aspis et.al., in [90] illustrates the generation of clusters using

trained perceptions, which are subsequently labeled using symbolic knowledge. Similarly, researchers in [91] overcome the limitations of deep neural networks, such as lengthy convergence times and overfitting data, by adopting the neuro-symbolic approach on big data. In this approach, big data is first transformed into a symbolic model, followed by embedding to create a training set. Other studies, such as [92], explore KiL in deep learning models, where the infusion of external representational knowledge from knowledge graphs assists in supervising feature learning and enhancing the model learning process.

3.5 PICO Framework

Not much work has been done in area of PICO based retrieval of contents from biomedical literature. The existing studies are categorized based on three aspects: individual PICO element identification [93-95], sentence classification [96][97], and Q/A & summarization [93,98]. In a proceeding, authors present a process of PICO corpus at individual element level and sentence level [99]. A hybrid approach of combining machine learning and rule base methods is proposed for identification of PICO sentence and individual elements in successive order [100]. A different work on PICO sentence extraction is carried out with Supervised Distance Supervision (SDS) approach that capitalizes on a small labelled dataset to mitigate noise in distantly derived annotations [101]. The authors developed a Naïve Bayes based classifier and reported that it is not sufficient to rely only on the first sentence of each section, particularly, when a high recall is required [95]. Multiple supervised classification algorithms were used to detect PICO elements at the sentence level in [102] by training data on structured medical

abstracts for each PICO element. Results show that the detection accuracy is better for P compared to I and O elements.

Based on above literature review, it is evident that current methodologies for curating and generating knowledge graphs (KGs) are subject to the following limitations:

a. KGs do not have the ability to handle the continual influx of new data sources and knowledge and lack the ability to continuously update contextual and temporal information.
b. Domain specific KGs cannot be generalized and have weaknesses in terms of concept extraction, clustering, relationship extraction, quality, and completeness.
c. The current KGs are not PICO-specific; having a PICO-enabled subgraph generated utilizing Neuro-Symbolic clustering techniques will make searching and indexing in KGs easier and contextualized.

CHAPTER FOUR

MATERIALS AND METHODS

This chapter provides a comprehensive overview of the methodology employed to address the research questions under investigation and discusses evaluation metrics employed to assess the effectiveness and performance of the overall framework. The chapter begins with a discussion of the dataset, followed by Neuro-Symbolic clustering framework and extraction of PICO elements. Next, the chapter presents a detailed overview of the data preprocessing techniques used to clean, transform, and integrate the data, followed by machine learning methods applied to the data, including clustering, topic modeling, and relation extraction. The methodology also includes the application of deep learning models such as BERT and BioBERT for embedding the PICO elements for evidence-based decision support.

4.1 Dataset

In the current study, the curation framework for knowledge graphs has been implemented and tested on two data domains specific to the field of health care, COVID-19 and subarachnoid hemorrhage. In addition, it uses relationships extraction datasets to extract triplets and identify relationships between concepts. The framework also uses generic and specific ontologies from BioPortal to extract hierarchical relationships per concepts.

4.1.1 COVID-19 Dataset

In the current framework, CORD-19 [103] dataset is being used to gather research papers that are related to COVID 19 and contain information about the COVID 19. The

CORD-19 database contains more than a million scholarly articles on COVID-19, SARS-CoV-2, and other coronaviruses. In the framework, the CORD-19 dataset was employed as the initial data source for collecting textual information on COVID-19. The abstract text of the papers was utilized, supplemented by the full paper text whenever accessible. Subsequently, a specific parsing framework, elaborated in section 4.2, was applied to categorize the sentences into PICO with Aim category. This dataset along with BioBERT encoded representations is publicly available to consume for further research in GitHub repository [104].

4.1.2 Cerebral Aneurysm Dataset

In order to assess the robustness and scalability of the framework on a different dataset, a PICO classification dataset was curated from biomedical literature. This dataset was developed based on a study of the research papers' quality, employing ensemble machine learning models and deep neural networks for PICO classification. Section 4.3 delineates the specifics of the PICO extraction framework. This dataset contains more than 170,000 abstracts related to subarachnoid hemorrhage and classified into A- Aim, P-Patient/Problem, C-Comparison, O-Outcome. **Table 3** shows the frequency per each PICO category.

4.1.3 Relationship Extraction Dataset

BioRel [105] is an imbalanced dataset for biomedical relation extraction offering reasonable performance as a starting point. The first step in this dataset was to identify Medline sentences mentioning entities, followed by MetaMap[106] references linked to UMLS (Unified Medical Language System).

Table 3: PICO elements distribution frequencies in aneurysm dataset

PICO Category	PICO Elements Frequency
A	31,676
P	49,308
I	11,711
R	30,515
O	50,072

In the end, each pair of entities in the sentence has its own relation label. A Multi-Instance-Learning (MIL) dataset, BioRel is divided into train, development, and test sets. Each set consists of n bags $\{\langle h_1, t_1, r_1\rangle, \langle h_2, t_2, r_2\rangle, \ldots, \langle h_n, t_n, r_n\rangle\}$, where h_i, t_i, and r_i are head entity, tail entity, and relation type. Each sentence consists of a sequence of k words $\{w_1, w_2, \ldots w_k\}$. The RE task focused solely on utilizing head entities (h), tail entities (t), and relation types. For example:

*Three cases of **acute myeloid leukemia** developing after treatment of renal disease with **cyclophosphamide** have been studied.*

Where 'acute myeloid leukemia' and 'cyclophosphamide' are head and tail entities. For this instance, in a bag, the relation type identified is **'treated by'**. The dataset contains 533,560 sentences, 26 million words, 69,513 entities, and 125 types of relationships. **Table 4** displays the distribution of the dataset across the training,

development, and test sets in terms of the occurrence of sentences, head entities, and tail entities.

Table 4: Distribution of sentences, head entities, and tail entities in train, development, and test splits for BioRel Dataset

Unique Count	Train	Development	Test
# Sentences	534,277	114,506	114,565
# Head Entities	36,961	16,550	16,583
# Tail Entities	37,607	16,902	17,011

4.2 Automated Knowledge Graph Curation Framework

The architecture of the automated deep-learning curated knowledge network is depicted in **Fig.2.** This framework utilizes Natural Language processing (NLP) techniques to cleanse text data, followed by a Neuro – Symbolic clustering approach to identify different topics in the article corpus. Subsequently, a classification model based on deep learning was employed to categorize the text into PICO elements. The classified text was then used in the KG modeling employing concept extraction using Ontology Based Information Extraction (OBIE) techniques, relationship extraction using BioBERT and CNN and Bidirectional LSTM (Bi-LSTM).

The proposed KG construction framework includes five major components: Data Preprocessing, Clustering, PICO classification, Relationship Extraction, and Knowledge Graph Generator. The main tasks of the data processing module are cleaning, stop words

removal, tokenization, and generating vector representations. The clustering module includes generation, identification, and optimization of clusters in the text corpus by using semantic enrichment, vector embedding and dimensionality reduction techniques. The PICO classification modules classified each paragraph into PICO elements and an additional category Aim by using RNNs based LSTM classifiers. The relationship extraction module utilizes generic and specific ontologies to extract hierarchical relationships and uses deep learning-based models to extract non-taxonomic relationships between concepts. Lastly, the knowledge generator module combines triplets, hierarchical relationships, and creates initial KGs that are optimized to remove ambiguous relationships and add missing links.

4.2.1 Data preprocessing

Data preprocessing is a crucial step in NLP that involves cleaning and transforming raw text data before it can be used for analysis or machine learning tasks. It plays a vital role in improving the quality and efficiency of NLP models and allow machine learning models to focus on relevant information handle challenges and achieve better performance. The COVID-19 dataset, as well as the aneurysm dataset contain texts that come from a variety of data sources. Thus, this step of preprocessing eliminates the inconsistencies and duplicates in the data, performs the text cleaning for NLP, and vectorizes the text into an embeddable format as well. The details of data processing steps such as text cleaning, concept extraction, tokenization and creation of embedding vectors are described as follows.

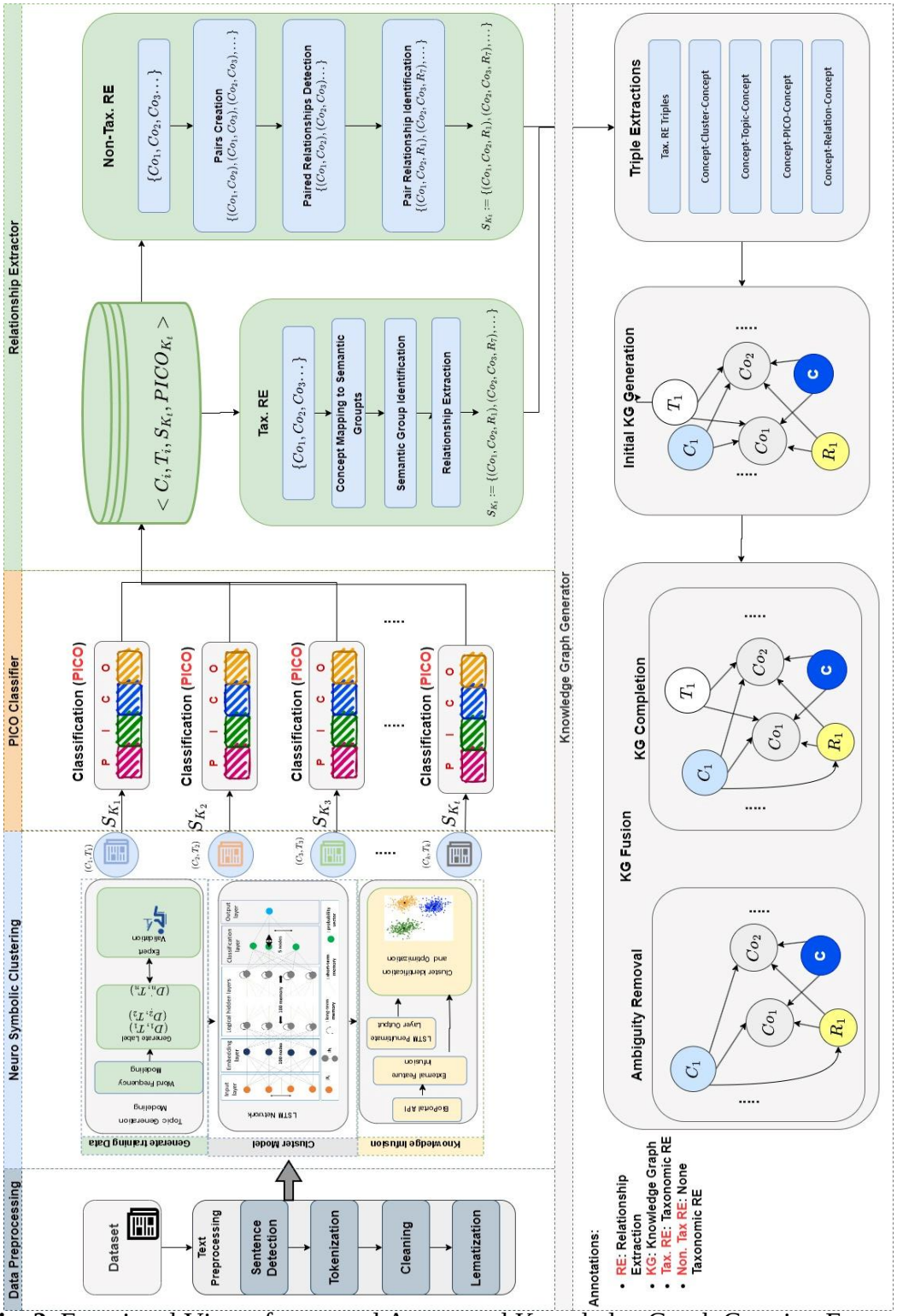

Fig. 2. Functional View of proposed Automated Knowledge Graph Curation Framework with five major components; Data Preprocessing, Clustering, PICO classification, Relationship Extraction, and Knowledge Graph Generator

4.2.1.1 Text Cleaning

In the text cleaning process, several techniques were employed to process the full-text research articles. These techniques included the removal of special characters, HTML tags, references, tables, and images. Following this, sentences were identified and punctuation and stop words were removed from the text using the Python Natural Language Toolkit (NLTK) library [107], which was augmented with medical-specific stop words. To identify special characters, symbols, and URLs, regular expressions were utilized. Once these text cleaning steps were completed, the final output was deemed to be cleaned and ready for tokenization.

4.2.1.2 Concept Extraction, Tokenization and Embedding Layer

The next step in our methodology involved the tokenization of sentences from the research articles, thereby dividing them into individual tokens. To extract concepts from the texts, the NLTK library was utilized to generate relevant concepts from the research papers. Additionally, experimentation was conducted using the BioPortal API [108] annotator method to extract concepts from the texts. This approach involved leveraging underlying generic and specific ontologies. The extracted concepts from each text were saved for subsequent enrichment.

To facilitate the use of machine learning models, an in-built tokenizer embedded in large language models (LLMs) like BERT and BioBERT was utilized. This process involved generating an embedding matrix with 768 dimensions.

4.2.2 Neuro – Symbolic Clustering

The purpose of clustering is to group similar research articles and extract related knowledge under one node in the KGs to make it more coherent. As the initial corpus of

research articles about coronavirus and subarachnoid hemorrhage is very large, a wide variety of topics are included within the articles. These topics comprised symptoms and diseases, corona vaccines, genetic information, risk factors, treatment methods, shapes of aneurysms, and genetic information and genetic screening. In order to make KGs better structured and easier to search for by following the different nodes specific to each topic, it was necessary to cluster these documents automatically. This clustering framework utilizes the preprocessed and tokenized texts generated in the data preprocessing steps and comprises of three key components: Training Data Generation, Cluster Model, and Knowledge Infusion. The training data generation involves topic modeling utilizing word frequency approach and generating labeled dataset with expert input. The model training module includes creating sequence to sequence models based on the labeled data. Lastly, the knowledge infusion module takes the penultimate layer output and optimizes the cluster by infusion of external knowledge from specific ontologies for identified topics.

4.2.2.1 Generate Training Dataset

In this module, the concepts derived from each document were utilized to generate word frequency metrics for each research article. By associating the concepts with the articles, different types of documents could be identified, including those pertaining to drugs, vaccines, symptoms, and genetics. To enhance the precision of the word frequency metrics, the top 100 most frequently encountered concepts were chosen from each article. This selection process yielded a concise and informative representation of the article's content. Subsequently, expert evaluation was employed to label the data into four distinct categories: drugs, vaccines, symptoms, and genetics. Through this approach, the unstructured text data was transformed into labeled data.

This approach allowed us to generate a labeled dataset that could be used for training machine learning models to classify research articles based on their content. The resulting dataset, enriched with expert labeling, provided a more accurate and targeted approach for analyzing the vast amount of research on COVID-19 available in the literature.

4.2.2.2 Cluster Model

The aim of this model is to develop a clustering model, which can understand the textual context and classify the document into four identified categories: Drug, Vaccine, Symptom and Genetics. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that has shown great success in sequence modeling tasks such as natural language processing. In this experiment, BioBERT Embeddings [109] were employed to vectorize the text data. BioBERT is a specialized language model that underwent pre-training on a vast collection of biomedical text. Extensive evaluations have demonstrated BioBERT's superiority over general-purpose language models for numerous biomedical natural language processing tasks.

The LSTM architecture consists of five main layers: the input layer, the embedding layer, the hidden layer, the classification layer, and the output layer. The embedding layer is responsible for transforming the input sequence into a 100-dimensional representation that captures the semantic meaning of each document. The hidden layer utilizes 100 memory units and a recurrent dropout rate of 0.2 to capture temporal dependencies in the sequence. The output layer is a dense layer that employs a SoftMax activation function to produce four nodes, corresponding to the four categories for which the model is trained. The categorical cross-entropy loss function

is utilized to optimize the model's performance during training. The model was trained using a batch size of 64 and trained for 5 epochs. To assess the model's performance during training, a validation set of 100 randomly selected documents was utilized.

4.2.2.3 Knowledge Infusion

The domain specific knowledge using external data sources can enable explanation, reasoning by creating new hypotheses and optimizing clustering. Biomedical repositories such as those in the field of healthcare and life sciences contain valuable information that is highly relevant to the identified clusters. For this experiment, BioBERT Embeddings were employed, and topic-specific information was extracted using domain-specific ontologies via the BioPortal API [108]. The knowledge infusion process consisted of three steps. Initially, the concepts matched to specific ontologies were extracted for each research article, and the normalized ratio of matched concepts to total concepts per article was calculated. In the subsequent step, synonyms and prefLabels were retrieved for these concepts, and embeddings were generated using BioBERT. Lastly, the obtained features were combined with the LSTM penultimate layer output, and this additional knowledge was infused into the model. The embedding of this additional knowledge created a new hypothesis on top of the clustering done using LSTM penultimate layer output and optimized the output of initial clustering results. **Table 5** presents the specific ontologies utilized for extracting knowledge and creating features, along with their respective types and names. **Algorithm 1** provides a comprehensive overview of the entire knowledge infusion process, outlining the step-by-step procedure for incorporating the additional knowledge into the system.

Algorithm 1 Knowledge Infusion Algorithm

Input: Documents [$d_1 \ldots d_{1000}$], LSTM penultimate layer output [$d_1 \ldots d_{1000}$]

Output: Optimized Cluster [1,2,3,4]

1: Let d= 1.

2: **while** d<=1000 **do**

while ontology in *[Drug, Vaccine. Symptom, Genetics]*

3: Fetch concepts from ontology

4: Calculate concept match

5: Normalize concept percent match

6: **Concept Match Pct**= Normalized concept match

7: Fetch Synonyms, PrefLabel from ontologies

8: Create semantic embedding using BioBERT

9: **Semantic Embedding**=BioBERT Embedding

10: Concatenate Features = [LSTM penultimate layer, Concept Match pct, Semantic Embedding]

11: **end while**

12: **end while**

13: **return** Concatenated Features

Table 5: Ontologies for external knowledge infusion

	Ontologies
Drug (D)	Drug Ontology (DRON)
	Prescription of Drugs Ontology (PDRO)
Vaccines (V)	Vaccine Ontology (VO)
	Vaccine Investigation Ontology (VIO)
	Vaccine Informed Consent Ontology (VICO)
Symptom (S)	Symptom Ontology (SYMP)
	Clinical Signs and Symptom Ontology (CSSO)
Genetic (G)	Gene Ontology
	Gene Ontology Extension

4.2.3 PICO Elements Extraction Framework

The purpose of the PICO classifier is to loop through each of the topics and categorize the texts into the various PICO elements along with any additional Aim categories that may be present. This categorization creates a subgraph specific to each

element, increases its relevance, helps in indexing, and improves the response time to query. The PICO classification is done for each identified topic in the cluster identification stage. To achieve this, the topic and its corresponding vectorized corpus input were segregated and fed into the RNN-based PICO classifier model as inputs. A separate framework for PICO Elements extraction was developed and then integrated into the knowledge graph curation framework.

The architecture of the PICO Elements Extraction Framework is depicted in **Fig.3.** This framework comprises four main modules, namely data preprocessing, quality recognition, PICO Classifier.

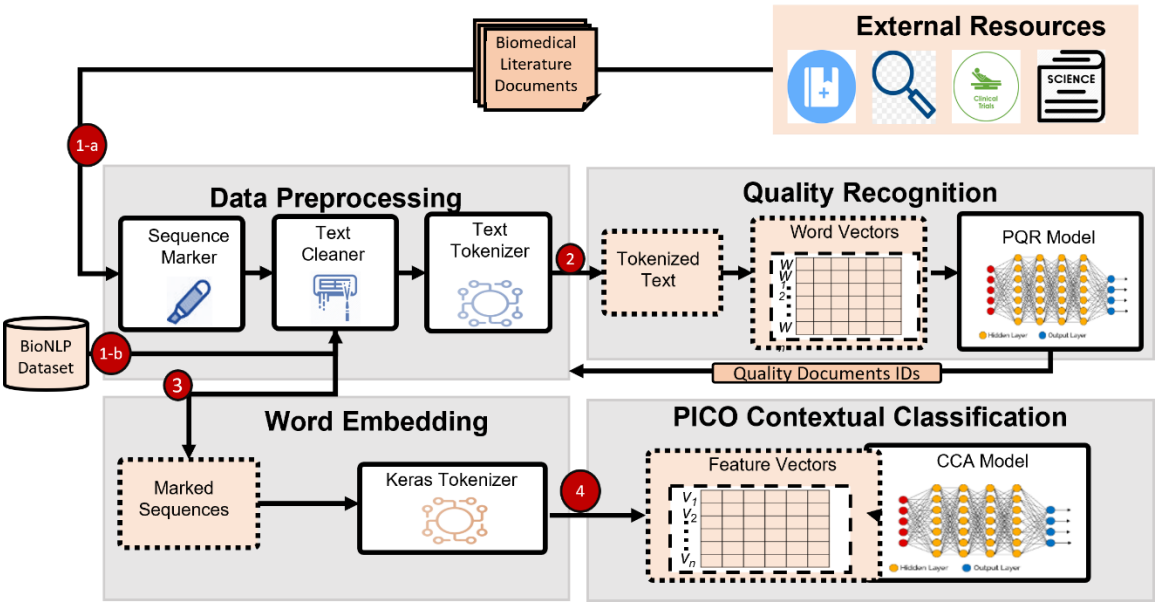

Fig. 3. Functional View of PICO Classification Framework with major components; data preprocessing, quality recognition, and PICO classification.

The data preprocessing module is responsible for pre-processing the raw biomedical data to remove any noise and irrelevant information. The main tasks in preprocessing steps includes sequence marker, cleaning, tokenization, vector generation, and feature creation. The quality recognition module of the proposed Biomed-

Summarizer framework employs a feed-forward deep learning binary classifier to identify the scientific soundness of studies based on biomedical data and metadata features. The classifier is trained on a large dataset of biomedical information to accurately determine whether the study is suitable for further processing. The CCA Classifier module, on the other hand, utilizes a Bidirectional Long Short-Term Memory (BI-LSTM) recurrent neural network to identify the PICO sequence of sentences in the selected quality documents. The BI-LSTM is trained on a vast dataset comprising domain-specific Medline abstracts and data acquired from BioNLP (Natural Language Processing for the biomedical domain). **Table 6** presents a sample text sequence extracted from an abstract, accompanied by the assigned class labels.

4.2.3.1 Preprocessing and Sequence Marking

A sequence marking technique was devised to identify PICO elements within the abstracts of the research articles. The role of the sequence marker is to parse each raw abstract in the dataset and extract the headings by matching the keywords listed in the dictionary, as illustrated in **Table 7**. When a keyword in the abstract is matched, the corresponding text (a sequence of sentences) under that heading is extracted and labeled accordingly. For example, if a heading like 'objective' is identified in an abstract, the text is assigned a label 'A'. This process is repeated for all the abstracts in the dataset of documents retrieved from PubMed®. After the completion of sequence marking, the subsequent steps involved text cleaning and concept extraction, as outlined in sections 4.2.1.a and 4.2.1.2.

Table 6: Example sequence of sentences for an assigned category of Aim, Population, Methods, Interventions, Results, Conclusion, and Outcomes

Text Sequence from abstract	PICO Category
Between January 2010 and December 2015, a total of 130 consecutive patients at our institution with the A com aneurysms-86 ruptured and 44 unruptured-were included in this study. The ruptured group included 43 females (50%) and 43 males (50%) with the mean age of 56 years (range, 34-83 years). The unruptured group included 23 females (52%) and 21 makes (%) with the mean age of 62 years (range, 28-80 years). All patients underwent either digital subtraction angiography or 3-dimensional computed tomography angiography. The exclusion criteria for this study were the patients with fusiform, traumatic, or mycotic aneurysm. There were preexisting known risk factors, such as hypertension in 73 patients, who required antihypertensive medication; other risk factors included diabetes mellitus (9 patients), coronary heart disease (9 patients), previous cerebral stroke (18 patients), and end-stage renal disease (3 patients) in the ruptured group. In the unruptured group, 38 patients had hypertension, 4 had diabetes mellitus, 5 had coronary heart disease, 10 had a previous cerebral stroke, and 2 had end-stage renal disease.	P
Four intracranial aneurysms cases were selected for this study. Using ct angiography images, the rapid prototyping process was completed using a polyjet technology machine. The size and morphology of the prototypes were compared to brain digital subtraction arteriography of the same patients.	M
After patients underwent dural puncture in the sitting position at L3-L4 or L4-L5, 0.5% hyperbaric bupivacaine was injected over two minutes: Group S7.5 received 1.5 mL, Group S5 received 1.0 mL, and Group S4 0.8 mL. After sitting for 10 minutes, patients were positioned for surgery.	I
The ruptured group consisted of 9 very small (<2 mm), 38 small (2-4 mm), 32 medium (4-10 mm), and 7 large (>; 10 mm) aneurysms; the unruptured group consisted of 2 very small, 16 small, 25 medium, and one large aneurysms. There were 73 ruptured aneurysms with small necks and 13 with wide necks (neck size >; 4 mm), and 34 unruptured aneurysms with small necks and 10 with wide necks.	R
The method which we develop here could become surgical planning for intracranial aneurysm treatment in the clinical workflow.	C
The prevailing view is that larger aneurysms have a greater risk of rupture. Predicting the risk of aneurysmal rupture, especially for aneurysms with a relatively small diameter, continues to be a topic of discourse. In fact, the majority of previous large-scale studies have used the maximum size of aneurysms as a predictor of aneurysm rupture.	O

To enable embedding, a maximum limit of 50,000 words per paragraph was set. Following this, a word embedding vector representation of 250 dimensions was generated using the Keras text-to-sequence [110] functionality, ensuring consistent vector length for each representation.

Table 7: Master dictionary representing keywords appear in headings in structured abstracts of biomedical studies. The dictionary is inherited from a study [21] with an extension of a few more keywords such as Patient 1 and Patient 2.

```
dict = {'A': ['objective', 'background', 'background and objectives', 'context',
'background and purpose',
            'purpose', 'importance', 'introduction', 'aim', 'rationale', 'goal', 'context',
        'hypothesis'],
        'P': ['population', 'participant', 'sample', 'subject', 'patient', 'patient 1','patient
        2'],
         'I': ['intervention', 'diagnosis'],
        'O': ['outcome', 'measure', 'variable', 'assessment'],
        M': ['method', 'setting', 'design', 'material', 'procedure', 'process',
        'methodology'],
        'R': ['result', 'finding'],
        'C': ['conclusion', 'implication', 'discussion', 'interpretation table']
       }
```

4.2.3.2 Quality Recognition Model

The proposed model, as shown in **Fig.4.**, comprises of multi-layer feed-forwarded neural networks so-called multi-layer perceptron (MLP). The MLP is further optimized with an ensemble method using AdaBoost. The final AdaBoost-MLP, called PQR (prognosis quality recognition) model, was trained on a dataset acquired automatically through PubMed® searches based on following two criteria: by selecting the 'Clinical Query prognosis' filter and choosing scope as 'narrow'. To construct and assess the model, the following steps were carried out: (a) Dataset preparation for training the deep learning models. (b)Training and fine-tuning of the deep learning model. Furthermore, a comparison between the proposed model and shallow machine learning models was

conducted, focusing on precision and recall metrics. To train and test the PQR model, a dataset comprising a total of 2,686 Medline abstracts was collected. Within this dataset, 697 abstracts were identified as positive studies, indicating their scientific soundness. The retrieval of positive studies was accomplished using the interface of Clinical Queries [102].

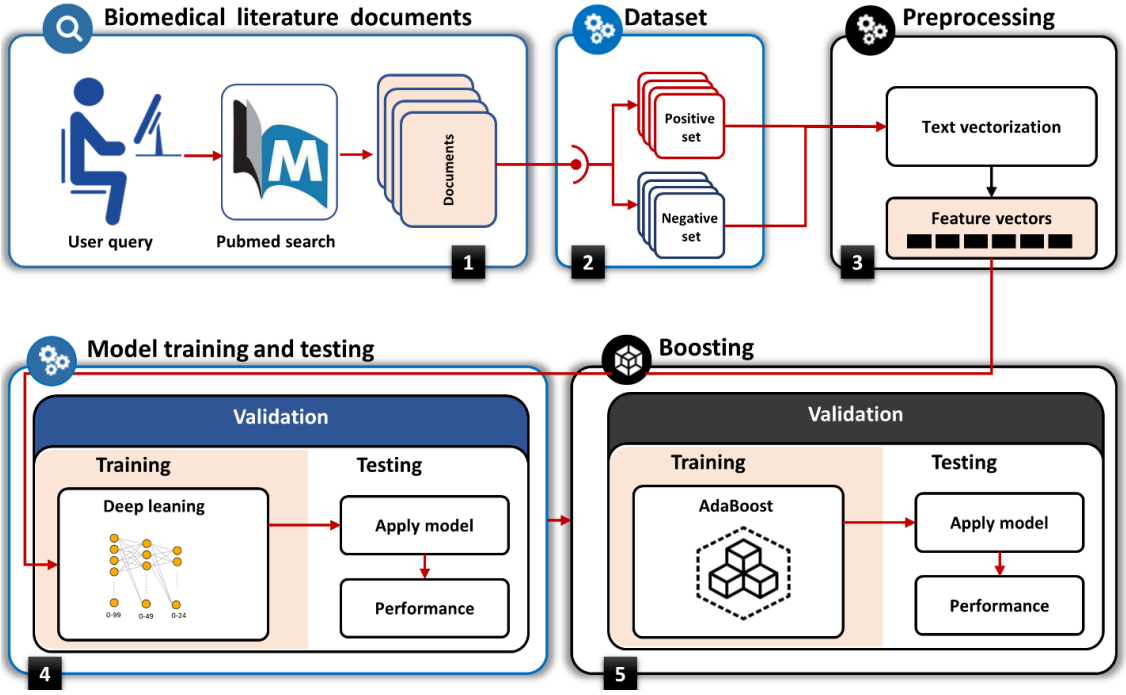

Fig. 4. Process steps of quality recognition model

4.2.3.2.1 Training Deep Learning Model

The PQR deep learning model is trained using five textual features, which include two data features (title and abstract) and three metadata features (article type, publishing journal, and authors). Prior to training, the data features undergo preprocessing steps outlined in the earlier section on data preprocessing. For the recognition of scientifically sound studies, a multilayer feedforward neural network model is employed for text classification.

The model consists of 5 layers with three hidden layers of size 100, 50, 25, respectively. The input layers take the BoW vectors generated from data features. The 'Maxout' activation function is utilized with 10 epochs during the training process. The data is split into a 70:30 ratio, with 70% allocated for training and 30% for testing purposes.

4.2.3.3 PICO Classification

The objective of this model is to create a multiclass classifier capable of accurately categorizing texts from chosen quality documents (determined by the quality model) into one of seven classes: A, M, P, I, R, O, and C. Subsequently, the M and P classes are merged into P due to their strong correlation observed in the actual identified texts associated with these categories. Similarly, the O and C classes are combined into C. While the focus primarily revolves around utilizing PICO classes in evidence-based medicine, the inclusion of other classes allows for non-PICO clinical applications in medical education and the development of a knowledge base for clinical decision support systems. PICO is a well-known terminology in evidence-based medicine. PICO is a widely recognized term in evidence-based medicine. The classifier is referred to as the PICO classifier because it encompasses all the components of PICO, along with two additional classes: A (aim) and R (results).

The PICO detection is a sequential sentence classification task instead of a single sentence problem; therefore, the sequence of sentences is jointly predicted. In this way, a better context is captured from multiple sentences, thus it improves the classification accuracy of predicting the current sentence. To build and evaluate the model, following steps were performed (a) preparation of a dataset for training the deep learning model, (b)

training and tuning deep learning model, and (c) comparison with traditional machine learning models in terms of precision and recall.

4.2.3.3.1 Model Building

Neural network model is heavily used in text processing due to the ability of processing arbitrary length sequences. Recurrent Neural Networks (RNNs) are immensely popular in multi class text classification. To construct and evaluate the classification model, the Bidirectional Long-Short Term Memory (Bi-LSTM) model is utilized. The Bi-LSTM is a type of RNN that effectively captures the long-term dependencies within the texts. It is widely recognized as one of the most popular models for text classification tasks.

The model is composed of five layers - the input layer, the embedding layer, the logical hidden layer, the classification layer, and the output layer. In this approach, the hidden layer is based on LSTM. The LSTM contains 100 units, considering $X_t$as data that entered the memory cell for training, it has three gates to control the flow of information in the network. The forget gate f_t controls how much of the previous state can be retained. The f_t defined as follows:

$$f_t = \sigma\left(W^f x_t + V^f h_{t-1} + b_f\right) \quad (1)$$

The input gate i_t in LSTM cell determines whether to use the current input to update the information of the LSTM or not using the following formulation:

$$i_t = \sigma(W^i x_t + V^i h_{t-1} + b_i) \quad (2)$$

The output gate o_t determines which parts of the current cell state need to be outputted to the next layer for iteration via:

$$o_t = \sigma(W^o x_t + V^o h_{t-1} + b_o) \quad (3)$$

Where W, V, and b are the weight matrix and biases, respectively. Also, σ is the sigmoid activation function which is defined as follows to control the weight of the message passing through.

$$\sigma(X) = 1 / (1 + e^{-X}) \quad (4)$$

At the end the output h_t in each cell can be retained as:

$$c_t = f_t \otimes c_{t-1} + i_t \otimes tanh(W^c x_t + V^c h_{t-1} + b_c) \quad (5)$$

$$h_t = o_t \otimes tanh(c_t) \quad (6)$$

Where $\otimes$ is the dot product.

As shown in the **Fig.5.,** the input layer consists of 250 dimensions showing the numeric features created using Keras tokenizer; the embedding layer, the LSTM logical hidden layers, the classification layer, and the output layer. The brief specification of the trained PICO classification model summary is provided as below.

- The first hidden layer is embedding layer that uses 100-dimension vectors to represent each paragraph.
- The next layer is LSTM layer with 100 memory units and recurrent dropout rate was set to 0.2.

- The next is the dense layer with 5 nodes and for the multi classification, theSoftMax activation function is employed and *categorical_crossentropy* is used as loss function.
- The training data is partitioned into two sections, with 10% of the data allocated for category validation and minimizing the loss function. Therefore, 140,454 instances are utilized for training, while 15,606 instances are reserved for validation purposes.
- The model is trained across 5 epochs with mini-batch size of 64.

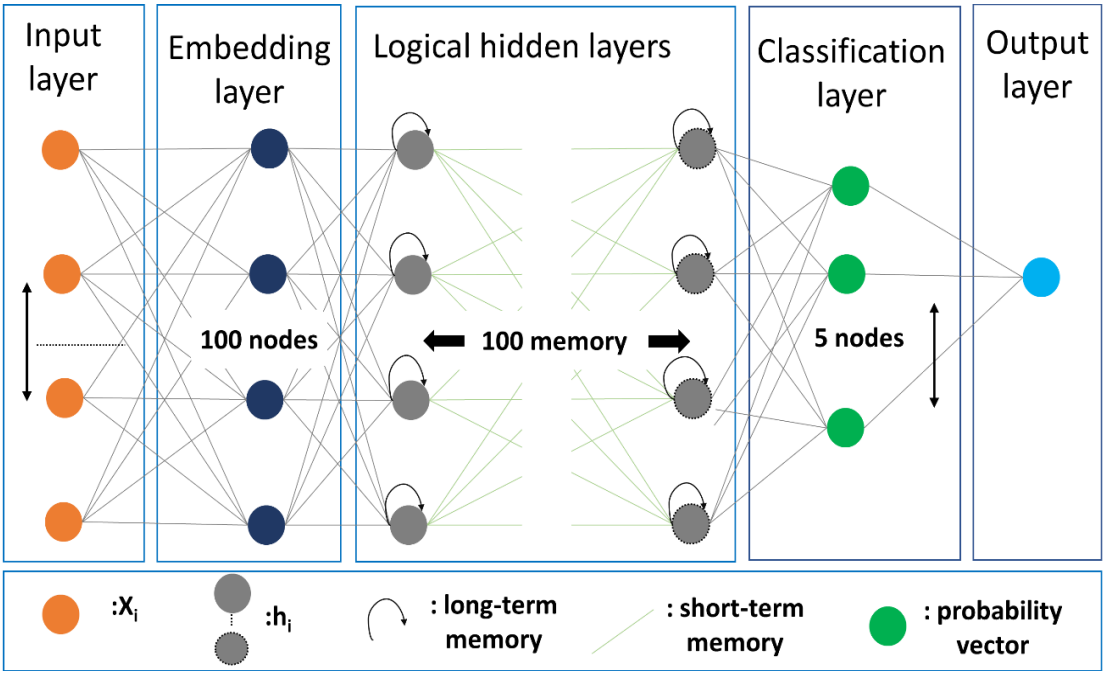

Fig.5.: PICO Classifier Neural Architecture

4.2.4 Relationship Extraction

This component was designed to identify taxonomic and non-taxonomic relationships between concepts that pertain to PICO elements pertaining to different topics. The detailed process of relationship extraction is given in the following subsections.

4.2.4.1 Taxonomic Relationship Extraction

The taxonomic relationship, also referred to as the "IS-A" relation, is one of the most important components of taxonomies, semantic hierarchies, and knowledge graphs. The Is-A relationship derived from specific and generic ontologies was utilized to establish connections among various concepts, merge them, and construct the knowledge graph (KG). Specifically, for the COVID-19 dataset, specific ontologies such as COVIDCRFRAPID [111], COVID-19[112], CODO [113], COVID19[114], IDO-COVID-19[115] were employed along with generic ontologies SNOMEDCT [116], MEDDRA [117]. In the case of aneurysm dataset SNOMEDCT, MEDDRA, NBO [118], NIFSTD [119] ontologies were utilized. The IS-A relationship is extracted by using maximum 3 levels up if available and constructing subgraphs in top-down fashion. The subgraph of concepts is further merged and joined through common entities. For example, one of the sentences in the aneurysm dataset is:

This Mendelian randomization study suggests that high blood pressure is a major risk factor for intracranial aneurysm.

The initial concept extraction contains two concepts *High Blood Pressure* and *intracranial aneurysm,* but after semantically enriching by extracting synonyms, definition and labels the final concepts identified are *High Blood Pressure*, intracranial *aneurysm*, *hypertension*, and *brain aneurysm*.**Fig.6.** shows the illustration of the merger of concepts based on taxonomic relation extraction and generation of complete subgraphs for the whole sentence.

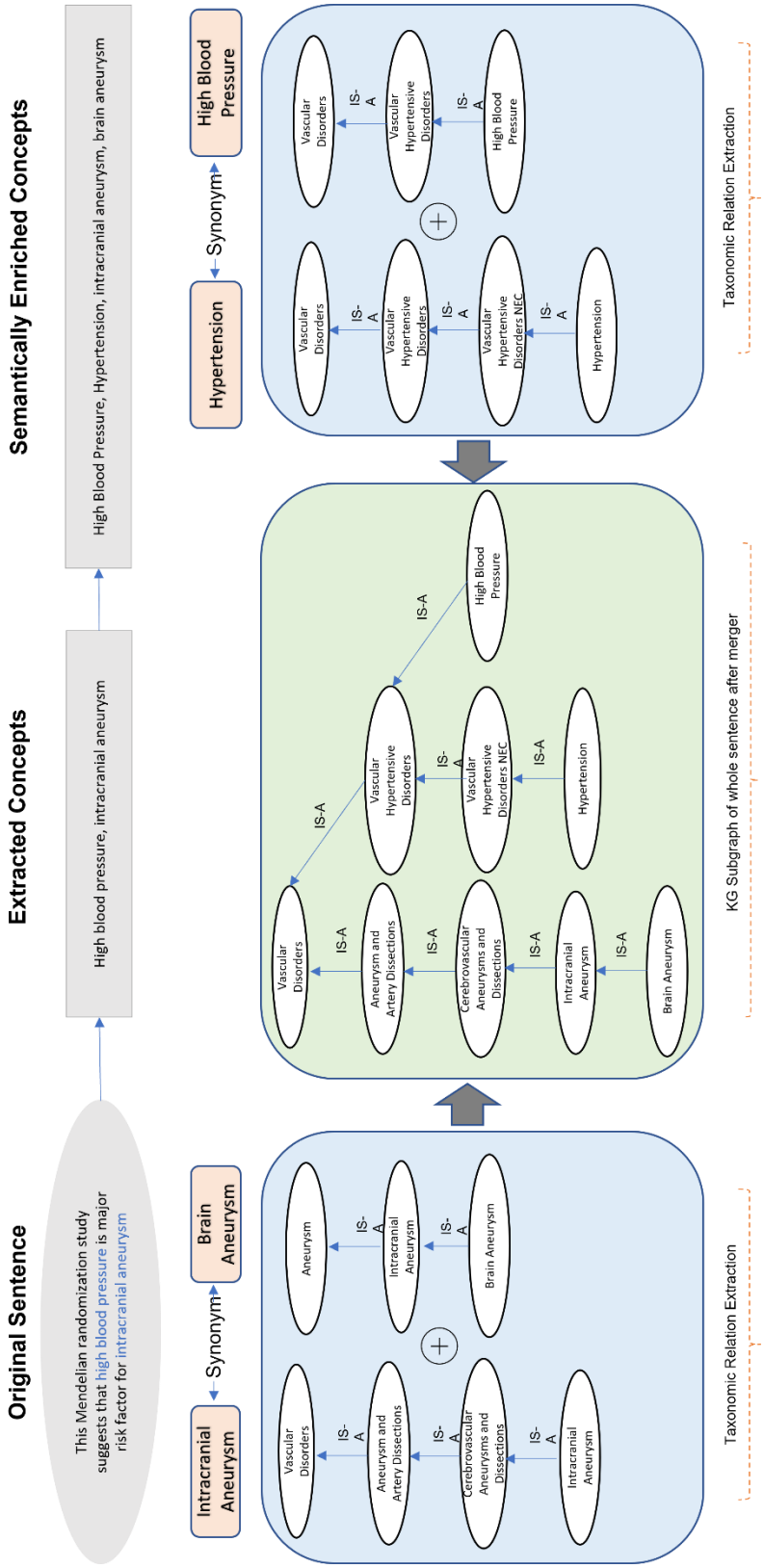

Fig. 6. Sample taxonomic relationship extraction and merger

4.2.4.2 Non-Taxonomic Relationship Extraction

A non-taxonomic relationship is a kind of 'part-of' relationship, in which one concept is a part of another. To enrich the KGs with non-taxonomic relationships, the main objective is to generate relationship triplets and to identify non-taxonomic relationships. Creation of non-taxonomic relationship between concepts is a two-step process, identifying the concepts with the relationship, followed by identification of the type of pairs with the relationship.

The non-taxonomic RE for KG generation formally defined as below:

Let's consider concepts {Co_1, Co_2, Co_3, …., Co_n} as inputs where n is the number of concepts in the list for a triple generation. The aim is to build a model f: (Co_i, Co_j) → argmax (R^r) where r is the dimension of the output relationships R and argmax is the function that outputs the relationship with maximum probability, i and j is the index of input concepts.

The following algorithm illustrates the main steps in constructing non-taxonomic relationships.

Algorithm 01: Pseudo code to extract non-taxonomic relationship

1. Inputs: {Co_1, Co_2, Co_3, …., Co_n}
2. Pairs creation: {(Co_1, Co_2), (Co_1, Co_3), (Co_1, Co_4), …. (Co_1, Co_n), …., (Co_{n-1}, Co_n)}
3. Paired relationships detection using model f, if argmax (R^r) < e (where e is the threshold for non-relationship pairs) then the probability for all relationship types is in the minority so the pair of (Co_i, Co_j) do not contain relationship: {(Co_1, Co_2), …. (Co_1, Co_n), …., (Co_{n-1}, Co_n)}
4. Pair relationship identification using model f: {(Co_1, Co_2, R_1), …. (Co_1, Co_n, R_7), …., (Co_{n-1}, Co_n, R_{125})}
5. Generate Output

In earlier approaches, feature-based and deep learning-based methods were tried, but the biggest problem was their representations and domains of use. The direct application of state-of-the-art NLP methodologies to biomedical text mining has some limitations. Although vector representations such as Word2Vec [120], GLoVe[121] by Stanford, and BERT have been examined in many general domains, it is difficult to estimate their performance on biomedical datasets. The word distributions of general and biomedical corpus also differ significantly, which can present problems for biomedical text mining. With the advancement of NLP and the high performance of the PrLMs, it was demonstrated that they could perform well in a specific domain such as medical texts because the attention mechanism enables them to consider context. One such model, BioBERT, which is a domain-specific language representation for the biomedical domain, has been taken into consideration as producing a representation layer to the RE extraction model.

The non-taxonomic relationship model includes deep learning and PrLMs using the following layers: representation layer, biomedical BERT layer, CNN layer, Bi-LSTM layer, and classifier for biomedical RE. The BioBERT model is based on the attention mechanism and transformer coding structure. BioBERT is used to obtain feature representations containing global semantic and contextual feature information of biomedical text, CNNs has a good performance on features extraction by convolution kernel, which improves the accuracy of feature descriptors. Here CNNs used to obtain features at different levels and combine attention to obtain weighted local features and fuse global contextual representation with weighted local features. This allows to Bi-LSTM layers to better capture the sequential representation of inputs for better

classifications. The detailed architecture for the model is shown in **Fig.7.** and components are described below.

4.2.4.2.1 Representation Layer

In the representation layer, words are split into their full forms or word pieces according to a set of rules, which are then encoded into numerical vectors. The representation consists of tokenization and concatenation of tokens. Each concept is tokenized using the BioBERT tokenizer. It produces the token, position, and segment embeddings.

$$Tokenizer(Co) = Co \rightarrow (W_k, P_k, S_k) \quad and \quad S_k, P_k, W_k \in \mathbb{R}^{200}, \quad k \in [1, 200] \tag{7}$$

Where, W, P, and S are token, position, and segment embeddings respectively. Next, embeddings are concatenated to produce the input sequence to the BioBERT module.

$$E_n = Tokenizer(Co_i) \oplus Tokenizer(Co_j) \ , \ where \ E_n \in \mathbb{R}^{400} \tag{8}$$

4.2.4.2.2 BioBERT Layer

Biomedical LMs are a natural choice as vectorization of inputs into semantic vectors since the proposed knowledge graph consumes biomedical text as input. A general-purpose language model, BERT achieves state-of-the-art performance in a wide range of tasks. The BioBERT is a biomedical variant of BERT that has been trained on medical datasets such as PubMed and PMC. Like BERT, BioBERT takes input sequence tokens and calculates embeddings for them. The framework employs the utilization of pre-trained BioBERT as the initial layer. Subsequently, fine-tuning is performed for the

RE task by incorporating feature extractor layers on top, allowing for the appropriate updating of the language model (LM) weights. The $V_{BioBert}$ represents the outputs of the BioBERT encoder which will be fed into the next layers for local feature extraction.

$$V_{BioBert} = Encoder_{12}\Big(Encoder_{11}(\ldots..(Encoder_1(E_n))\ldots.)\Big)$$

$$, \ where \ V_{BioBERT} \in \mathbb{R}^{768}$$

(9)

4.2.4.2.3 Convolutional Neural Network Layer

CNNs [122] are neural networks which extract local features. This process allows the extraction of high-level features and concept sharing. Thus, the network learns the importance of both concepts regarding relation types. The convolutional layer was followed by a ReLU (Rectified Linear Unit) activation function [123] on the extracted local features to produce the nonlinear representations. Ultimately, this layer applies a 1D convolution to an input composed of several input planes.

$$out(N_i, C_{out_j}) = bias(C) + \Sigma^{C_{in}-1}{}_{k=0} \ (weight(C_{out}, k) \star V_{BioBert}(N_i, k))$$

(10)

Where $\star$is the cross-correlation operator, N is the batch size, C is a number of channels (C_{in}, C_{out} are number of input and output channels respectively), L is a length of the signal sequence where it is equal to

$$L_{out} = \lfloor (L_{in} + 2 * padding - dilation * (kernelsize - 1) - 1)/stride + 1 \rfloor$$

(11)

Where padding is the same, dilation is 1 (the spacing between kernel elements), kernel size is 3, and the stride is 1. Next, a nonlinearity is applied to $out(N_i, C_{out_j})$ to capture the final local features.

$$CNN_{out} = ReLU\,(out(N_i, C_{out_j})) \qquad where \qquad ReLU(X) = Max\,(X, 0)$$

(12)

In the end, dropout randomly zeroes some of the elements of the inputs with probability p=0.3 using samples from a Bernoulli distribution.

4.2.4.2.4 Bidirectional LSTM

A bidirectional LSTM on top of a CNN layer eliminates the need for more feature engineering techniques by capturing local features and extracting them for long dependency features. On top of the CNN layer is a BiLSTM layer that is composed of two LSTM networks that can read input texts both forward and backward. The forward LSTM processes information from left to right ($H^{\rightarrow}$) and the backward LSTM processes information from right to left ($H^{\leftarrow}$), Next, the output of forward and backward LSTMs is combined and fed into the next layer. The following equation represents steps:

$$H_t{}^{\rightarrow} = LSTM\,(CNN_{out}\,, H_{(t-1)}{}^{\rightarrow})$$

(13)

$$H_t{}^{\leftarrow} = LSTM\,(CNN_{out}, H_{(t-1)}{}^{\leftarrow})$$

(14)

$$BiLSTM_{H_t} = [H_t{}^{\rightarrow} : H_t{}^{\leftarrow}] \qquad where \quad BiLSTM_{H_t} \in \mathbb{R}^{n*400}$$

(15)

4.2.4.2.5 Classifier

The CNN-Bi-LSTM layers contain enriched representations of input concepts; the classifier tries to learn these representations with respect to relation types. Backward propagation allows each layer of the network to update its weights. The classifier consists of hidden states h_t as inputs to the ReLU activation function followed by a dropout (p

=0.3) each neuron will be zeroed out independently on every forward call. Next, a linear transformation is applied to the hidden states, resulting in 125 output features from the initial 768 input features. Lastly, the outputs are fed into the argmax function to output the class with maximum probability as the relation type of the input concepts. The equation is described conceptually below, and source code of implementation is available on GitHub repository [124]

$$CNNBiLSTM_{out} = ReLU(BiLSTM_{H_t}) \quad (16)$$

$$Y := bias + CNNBiLSTM_{out} * weight^T \quad (17)$$

$$argmax\, Y(p) := argmax_{p \in Y}\, p = \{ p \in Y : Y(p) \geq Y(i) \ for\ all\ i\ in\ Y \} \quad (18)$$

4.3 Evaluation Metrics

The knowledge framework is evaluated on both Covid and aneurysm dataset. Different metrics have been used to evaluate each component of the automated knowledge framework. Neuro-Symbolic clustering, PICO classification models and relationship extractor models are evaluated and compared to baseline machine learning based classifiers using metrics accuracy, precision, recall, and F1 score. In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.

$$Accuracy = (TP+TN) / (TP+FP + TN + FN)$$

The precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.

$$Precision = TP/(TP + FP)$$

The recall is used to measure the fraction of positive patterns that are correctly classified.

$$Recall = TP/(TP + TN)$$

The F1 score metric represents the harmonic mean between recall and precision values.

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

Where TP, TN, TP, FN are defined as follows:

- True Positive (TP): Predicted *positive* and actual also *positive*
- True Negative (TN): Predicted *negative* and actual also *negative*
- False Positive (FP): Predicted *positive* but actual *negative*
- False Negative (FN): Predicted *negative* but actual is *positive*.

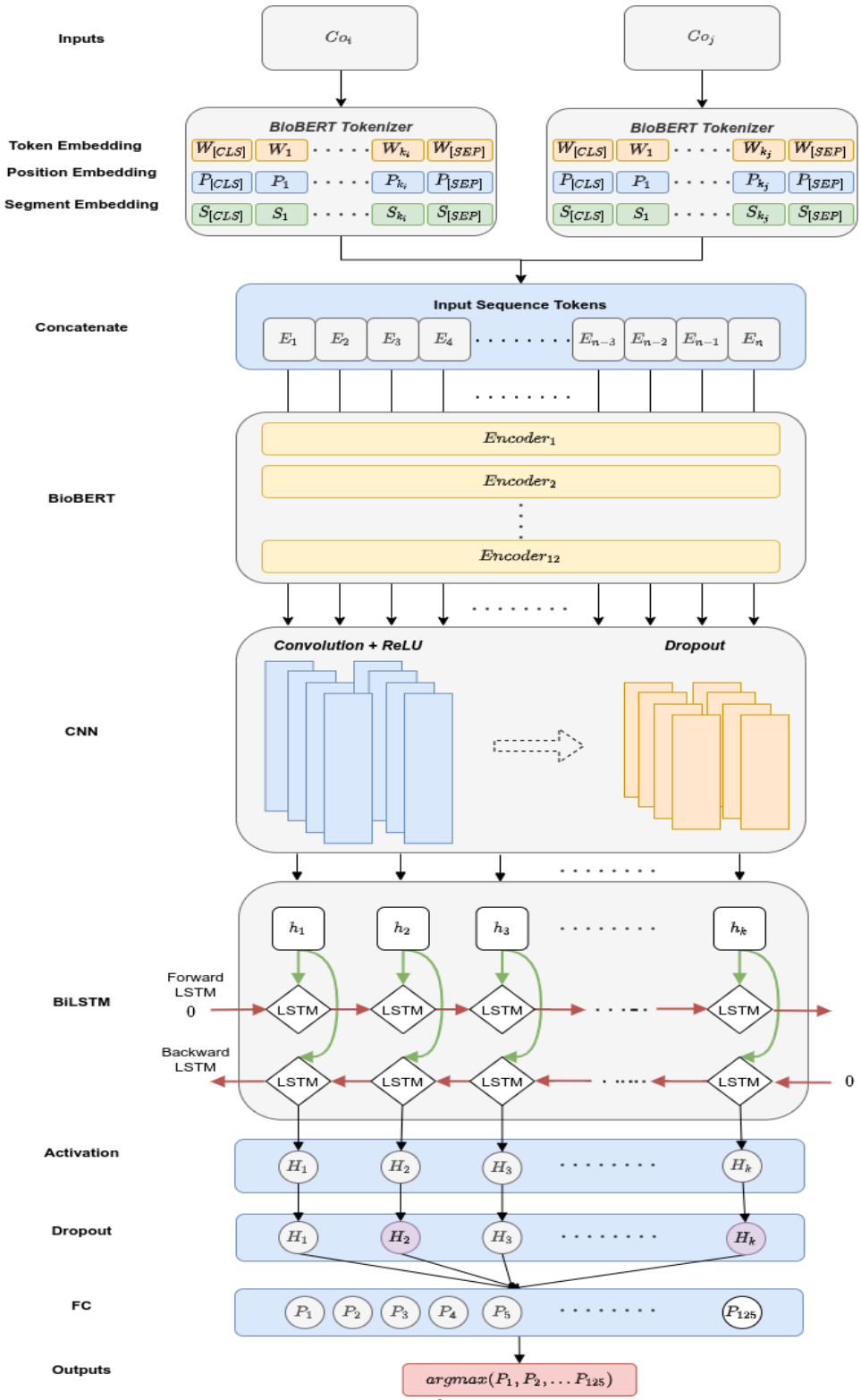

Fig. 7. Architecture diagram of relationship extraction module

CHAPTER FIVE

RESULTS

This chapter presents a comprehensive evaluation of the clustering, concept extraction, PICO classification, and relation extraction components. The evaluation process involved testing the proposed automated framework for knowledge graph curation on two healthcare datasets: cerebral aneurysm and COVID-19. The analysis of the results provided valuable insights into the benefits of the proposed framework for evidence-based decision support systems in healthcare.

5.1. Cluster Model

5.1.1 Dataset Preparation

To evaluate the proposed clustering framework, a randomized sample of 1000 research articles from the dataset was selected and labeled using topic frequency modeling and human effort.

5.1.2 Experimental Setup

To evaluate the clustering results, several machine learning models, including K-means, Zero Shot Learning (ZSL) [125], and RNNs such as LSTM, were employed. Furthermore, both generic embedding techniques like BERT and domain-specific embeddings such as BioBERT were utilized. Through these experiments, the effectiveness of various approaches was assessed, enabling the identification of the most suitable techniques for the specific domain.

5.1.2.1 K-Means Clustering

K-means clustering is a widely used unsupervised machine learning algorithm that partitions data into K clusters based on similarities in the data. One of the primary advantages of k-means clustering is its scalability to large datasets. Additionally, it is a simple and fast algorithm that can be applied to a wide range of applications such as image segmentation, customer segmentation, and anomaly detection. At first, BERT encoding was employed to generate vectorized input, which was subsequently subjected to PCA [126] to reduce dimensionality based on variance explanation. The elbow method [127] was then applied to determine the initial number of clusters, which was set to four (n=4). The dimensionally reduced vector and optimized number of clusters are fed to the K-means model to generate the cluster and later it is mapped back to original documents to specify the name of the clusters. **Fig. 8.** shows the clustering optimization result of COVID-19 and brain aneurysm dataset. For both the dataset, the sum of squared distances (inertia) falls suddenly at $k=4$, so the optimized number of clusters is determined as 4 and hence four topics are identified. After reducing the texts represented in vectors to two components, it is evident that there is a clear distinction between optimized cluster numbers.

5.1.2.2 Zero Shot Learning with BERT embeddings

BERT (Bidirectional Encoder Representations from Transformers) [31] is a pre-trained language model that has achieved state-of-the-art results on various NLP tasks. Zero-Shot learning (ZSL) is a type of machine learning technique that enables models to recognize and classify objects or concepts that have not been seen during training. This approach allows the model to generalize to new categories by learning a mapping

between different domains. One of the key advantages of zero-shot learning is its ability to overcome the limitations of supervised learning, where the model is trained only on labeled data. With zero-shot learning, models can be trained on a smaller set of labeled data and still generalize to new categories. ZSL was implemented utilizing a transformer architecture and embeddings generated by the BERT tokenizer.

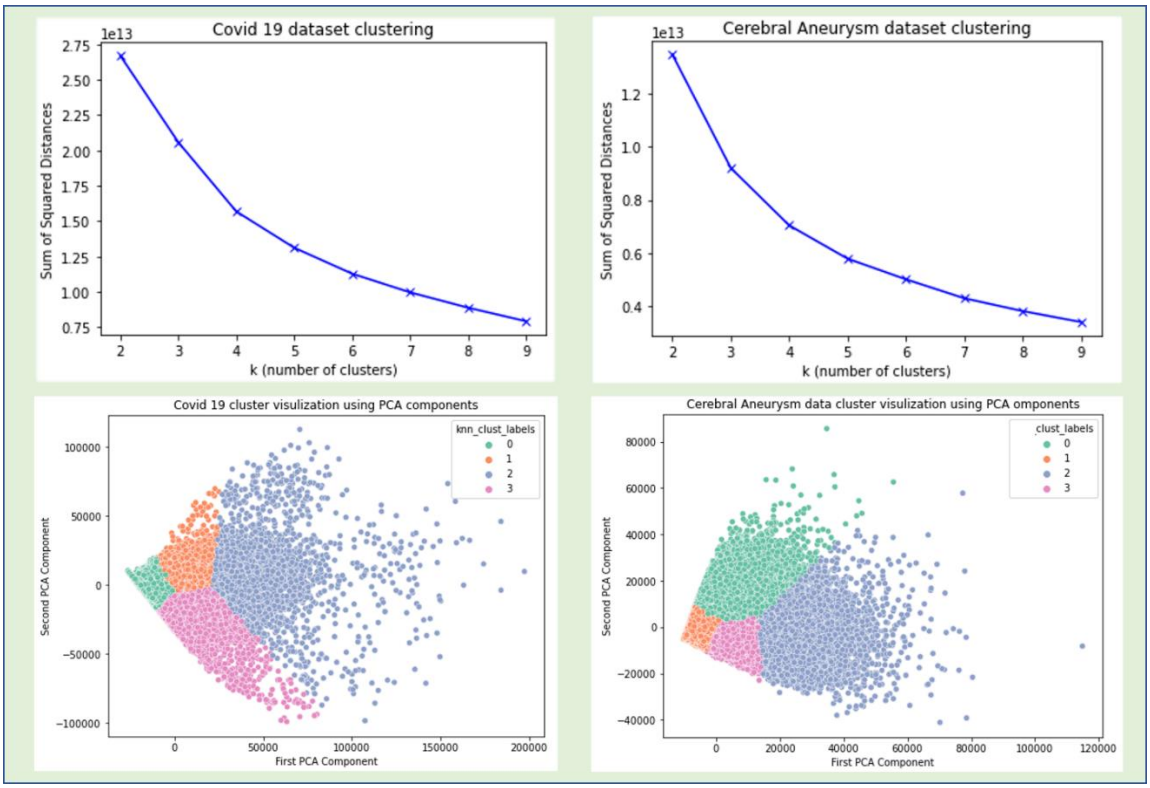

Fig. 8. Cluster optimization and selection using K-means clustering and elbow method

5.1.2.3 LSTM with BERT, BioBERT and Knowledge Infusion

In this approach, BERT and BioBERT embeddings are used to capture contextual information from the input text, which is then fed into an LSTM layer to capture the sequence information. The LSTM layer can then model the long-term dependencies in the sequence and make predictions based on the contextual and sequential information learned from the embeddings. The LSTM architecture utilized in the methods section was

also employed in this iteration. Subsequently, knowledge extraction from domain-specific ontologies was performed and integrated with the clustering results, resulting in the creation of new hypotheses and optimization of the cluster output.

Table 8 presents a comparative analysis of model performance across different categories using precision as evaluation metric. The results indicate that k-means clustering with BERT-encoded vectors had an average precision of 45%. Zero-shot learning with BERT vectors improved the precision by 22% due to the transformer architecture of ZSL, which can understand contextual information and hold important sequences using its attention layer. The LSTM model with BERT encoding further improved the performance as it was trained on the labeled dataset, unlike k-means clustering and ZSL, which had no labeled dataset. The LSTM model with BioBERT embeddings performed the best with an average precision of 76%, an improvement of 5% on the LSTM model with BERT. The specificity of embeddings in the biomedical domain is the reason behind its effectiveness, as it provides a deep contextualization of vector inputs with domain-specific concepts.

Finally, the best-performing model was the LSTM with BioBERT embedding with external knowledge infusion using specific biomedical ontologies, which had an average precision of 88%. This indicates that the external knowledge was able to aid in the existing clustering mechanism and improved the performance by 12%. **Table 9** presents a detailed performance evaluation using precision, recall, and F1-score.

5.2 Concept Extraction

The evaluation of the concept extraction module was conducted on a set of 100 randomly selected abstracts from the corpus. These abstracts were manually annotated,

and concepts were extracted for comparison with concepts extracted from the BioPortal API. The evaluation yielded a precision of 91%, recall of 78%, and an F1 score of 84%. The utilization of both generic and specific ontologies in both datasets contributed to achieving a high precision and ensuring the extraction of relevant terms without including irrelevant ones.

Table 8: Comparative result (Precision) of traditional model and LSTM with Knowledge Infusion across categories Drug(D), Vaccine(V), Symptom(S), Genetics(S)

Topics / ML Model	D	V	S	G	Avg
K-means Clustering					
BERT ZSL	0.47	0.42	0.53	0.37	0.45
BERT LSTM	0.69	0.63	0.67	0.68	0.67
BERT LSTM	0.72	0.74	0.62	0.77	0.71
BioBERT LSTM	0.73	0.69	0.79	0.82	0.76
BioBERT Knowledge Infusion	0.89	0.82	0.91	0.90	0.88

Table 9: Precision, Recall, F1-Score of proposed LSTM model with external knowledge infusion

Topics	Precision	Recall	F1-Score
Drug	0.89	0.81	0.84
Vaccines	0.82	0.81	0.81
Symptoms	0.91	0.87	0.84
Genetic	0.90	0.86	0.87

5.3 PICO Classification

The aim of this experiment was to quantify the comparative analysis of the proposed LSTM model with other baseline classifiers such as LR, KNN, Adaptive Boosting, Gradient Boosting, and Multi-Layer Perceptron (MLP) models on both the covid, and aneurysm dataset using keras embedding layer and BioBERT embedding layer

for initial input texts. The LR model was initialized with base configuration, KNN model was started with five neighbors and Euclidean as distance measure, Gradient Boosting model was initialized with learning rate =.10 and keeping rest of the parameters as default, and lastly MLP was initialized with hidden layer of sizes 100, 70, 30, 20 and 10 neurons, and maximum iteration of 1000. Results of these models are compared on four metrics: precision, recall, F1-score, and accuracy as shown in **Table 10** and **Table 11** on covid and aneurysm datasets. The proposed model is an RNN based LSMT model using BioBERT tokenization and embedding layer as an input with 82 % of precision, recall, accuracy, and F1-score on COVID-19 dataset and 93 % precision, recall, accuracy, and F1-score on aneurysm dataset.

5.4 Non-Taxonomic Relationship Extraction Evaluations

The aim of this experiment was to quantify the comparative analysis of the proposed RE model with other baseline machine learning model such as LR, MLP with Latent Semantic Analysis (LSA) encoding, BERT and BioBERT Model with encoding using softmax classifier.

5.4.1 Logistics Regression with Latent Semantic Analysis

This model uses LSA [128] as a representation layer and LR as a classifier. LSA embeds documents in a vector space using a bag of words method. This LSA is based on Term Frequency-Inverse Document Frequency (TF-IDF) and Singular Value Decomposition (SVD). The inputs were represented by 100 unigrams based on TF-IDF with SVD n-components of 50.

Table 10: Comparative results of deep learning with traditional machine learning models for different representation layer on COVID-19 dataset

Classifier	Embedding Layer	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Keras	38	37	33	42
	BioBERT	77	77	77	77
k-Nearest Neighbors	Keras	41	39	39	39
	BioBERT	75	74	74	74
Adaptive Boosting	Keras	37	45	39	45
	BioBERT	58	58	58	58
Gradient Boosting	Keras	48	48	42	48
	BioBERT	64	63	62	63
Multi-Layer Perceptron	Keras	8	28	12	28
	BioBERT	63	64	64	64
Long Short-Term Memory	Keras	72	71	71	71
	BioBERT	82	82	82	82

Table 11: Comparative results of deep learning with traditional machine learning models for different representation layer on aneurysm dataset

Classifier	Embedding Layer	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Keras	36	42	36	42
	BioBERT	76	74	75	77
k-Nearest Neighbors	Keras	35	35	34	35
	BioBERT	80	78	78	80
Adaptive Boosting	Keras	49	50	47	50
	BioBERT	64	66	66	66
Gradient Boosting	Keras	57	49	45	49
	BioBERT	64	61	59	61
Multi-Layer Perceptron	Keras	37	29	13	29
	BioBERT	83	83	83	85
Long Short-Term Memory	Keras	85	84	84	84
	BioBERT	93	93	93	93

5.4.2 Multi-Layer Perceptron with Latent Semantic Analysis

The model employs LSA as the representation layer and MLP as the classifier. The MLP classifier optimizes the log-loss function using the Adam optimizer. The architecture consists of three fully connected dense layers with 500, 300, and 250 neurons in each layer, respectively. The model is trained for 50 epochs with a batch size of 200 and a learning rate of 0.001.

5.4.3 Hypertuned BERT with Softmax Layer

A fine-tuned BERT in the usual multi-classification is one of the baselines in comparison to modified/integrated transformer models. A basic BERT-based model as a baseline for the RE task consists of the BERT, linear, dropout, and linear layers in sequential form (the BERT → Linear1 → Dropout → Linear2 scheme). A BERT outputs a CLS token as an input for the next linear layer, where it takes input features in size 768. Next, a dropout, a regularization method that approximates training a large number of neural networks with different architecture in parallel with a probability of 0.3 has been added. In the end, the second linear layer inputs feature in size 768 and outputs predictions in size of 125 (the number of unique relationships in the whole dataset). The argmax function is applied to the outputs of the second linear layer to the output relationship for input concepts. The training was done using the following parameters: a batch size of 4, Adam optimizer with a learning rate of 10e-05, 2 epoch numbers, and cross-entropy loss function.

5.4.4 BioBERT with Softmax Layer

The model has been trained following the same approach as presented in the fine-tuning of the BERT model, utilizing the following settings: BioBERT → Linear1 →

Dropout → Linear2. In the first layer instead of the BERT model BioBERT, a medical domain transformer model has been utilized.

The RE model is evaluated against above mentioned baseline models with varying embedding layers and results are shown in **Table 12.** The relation prediction classification is highly imbalanced, so traditional machine learning models like LR with LSA embedding are not able to capture high performance features. A simple neural network model with LSA embeddings performs better than LSA-LR and it shows that deep learning models are able to capture important features. Additionally, it was observed that domain-specific language models outperform general-purpose models like BERT.

The proposed CNN-Bi-LSTM model, utilizing BioBERT embedding, achieved a 2% higher F1-score compared to previous approaches. This improvement can be attributed to the ability of CNNs to learn higher-level features. Furthermore, traditional machine learning models do not perform as well as the proposed method since they are not able to capture high performance features.

5.5 Process flow: KG generation process for aneurysm treatment

Fig.9 illustrates the step-by-step process employed to generate the knowledge graph for aneurysm treatment. Initially, a cluster of 1309 documents is gathered from the PubMed data source using generic keywords related to cerebral aneurysm, including drugs, symptoms, and treatment. Subsequently, these documents undergo classification into prognosis (n=778) and non-prognosis (n=531) categories using a clinical interface query. In the third step, a quality recognition model is utilized to further classify these documents into positive (n=444) and negative studies (n=334). Moving to the fourth step, the positive studies' documents are inputted into a clustering model, resulting in the

generation of different topics such as drugs, symptoms, and treatment. Continuing to the fifth step, the treatment topic is processed through a PICO classification model, which categorizes sentences into P (n=53), I (n=34), C (n=59), and O (n=95) elements. Finally, by consuming the PICO elements, the knowledge graph is generated, and a subgraph representing the results is depicted in **Fig.9**. In the example subgraph for aneurysm treatment, intervention (I) for medium aneurysm at Anterior Communicating Artery (ACoA) is identified as endovascular treatment with 88.4% good outcome. The control measure i.e., comparison (C) determined in the KG is surgical clipping with 73.2 % positive outcome (O).

Table 12: Comparative results of proposed BioBERT encoded CNN based BiLSTM model with other baseline models

Classifier	Precision	Recall	F1-Score	Accuracy
LSA-LR	16	8	9	45
LSA-MLP	60	34	40	66
BERT	83	74	76	93
BioBERT	93	86	88	95
BioBERT-CNN-BiLSTM	92	90	90	96

5.6 Case Query: Covid Treatment Evidence

To test the effectiveness of the automated KG generation framework on the COVID-19 dataset, a simple query related to COVID-19 treatment was executed. In the

subgraph as shown in **Fig.10**., it searches through the entire KG and traverses through specific topic *treatment* that was identified through clustering and topic modeling in the framework and provide evidence relating to problem, interventions, comparisons, and outcomes. In the example subgraph for COVID-19 treatment, patients of age between 56 yrs. and ≤ 73 yrs. with diabetes and hypertension identified as problem(P), immunomodulators and anticoagulant are identified as intervention (I) with outcome as 2-3 weeks of hospital stay (O). The control measure (C) identified as use of antibiotic and antiviral medication with 3-4 weeks of hospital stay (O).

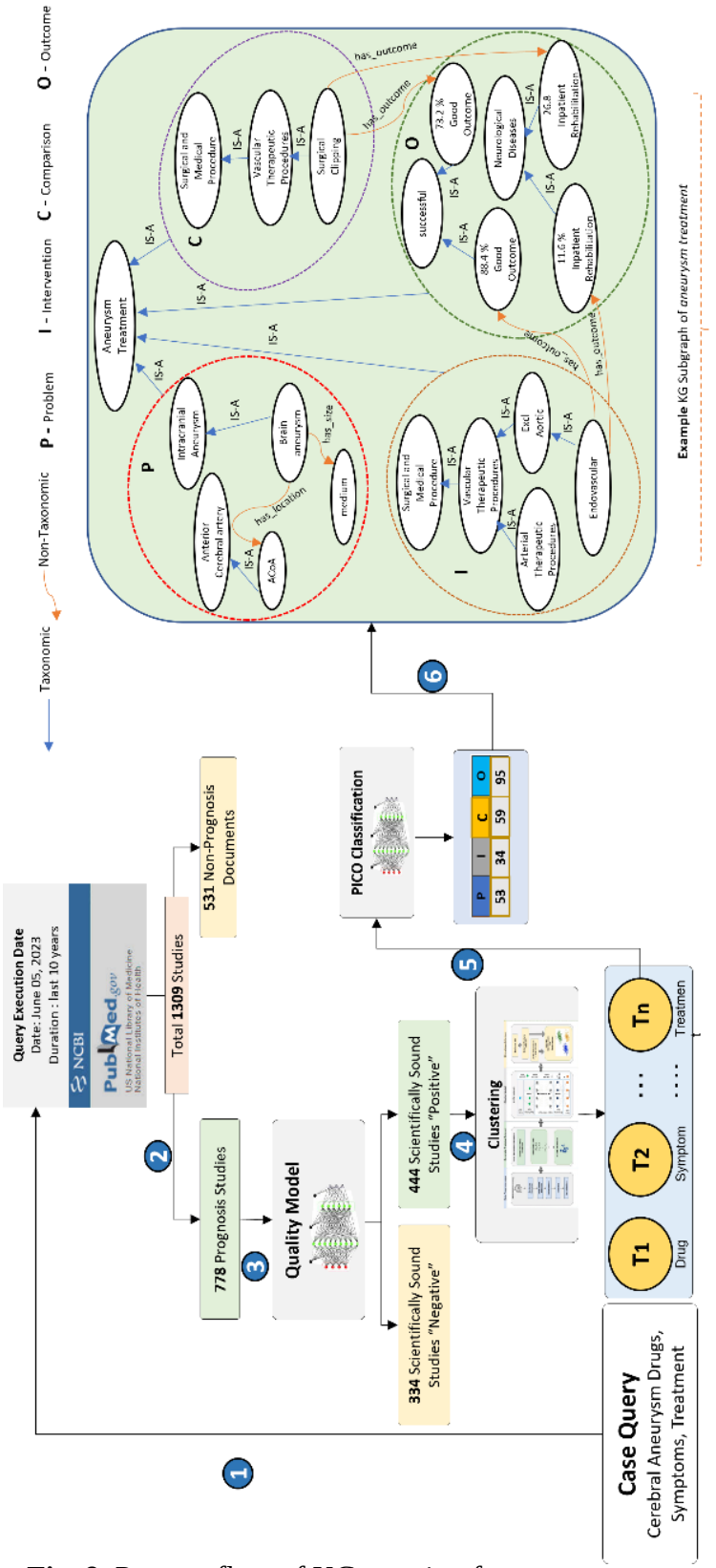

Fig. 9. Process flow of KG curation for aneurysm treatment

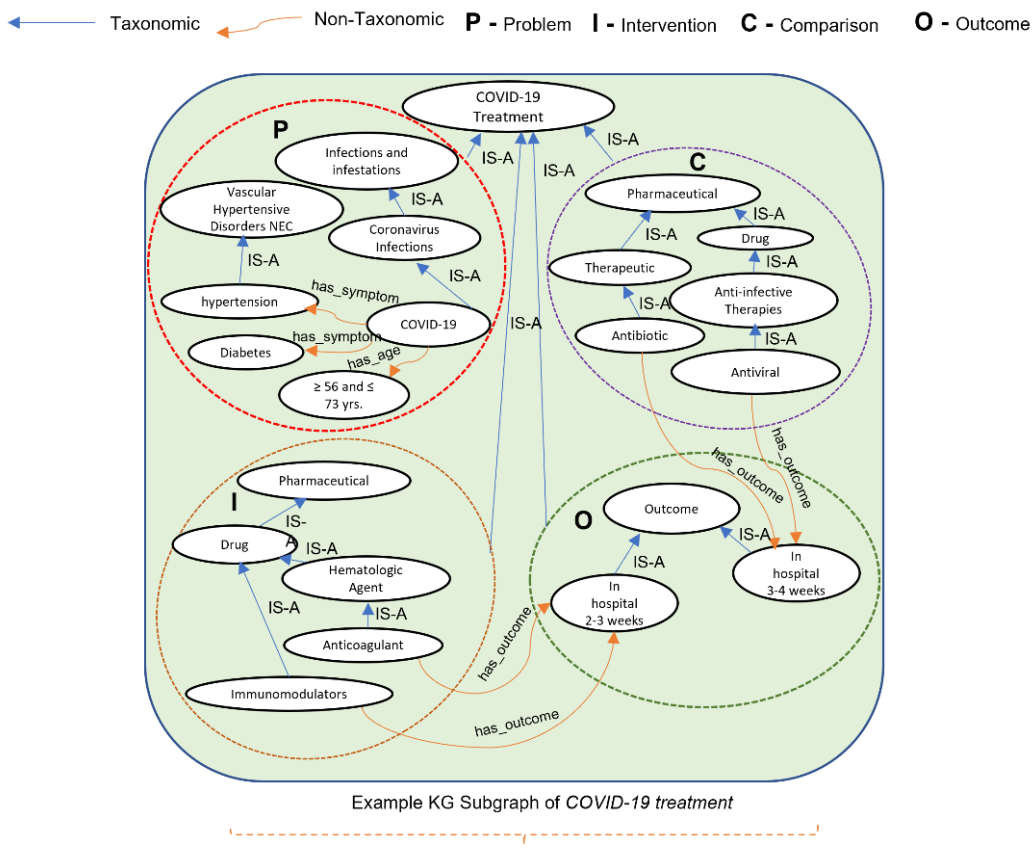

Fig. 10. An excerpt/sub-graph of aneurysm/covid treatment evidence

CHAPTER SIX

DISCUSSION, CONCLUSION AND FUTURE RESEARCH

Knowledge Graphs are expected to be an important technology to transform evidence-based medicine. As the underlying data source and inherent information keeps changing in literature, the automated knowledge graph generation framework for evidence-based medicine is unable to generate knowledge that is easy to query. The developed architecture of automated KG construction was shown to be flexible and adaptable to new sources of information by enabling automatic clustering of new information and assigning it to existing clusters or even creating new clusters followed by PICO classification which in turn makes deployment and storage of information easier and smarter. To the best of our knowledge, this is the first attempt to create a framework for PICO enabled KG generation that is generalizable to different clinical domains.

In KGs, responsiveness plays a crucial role in determining its overall performance. Su et al. [129] highlighted the significance of timely response, as it enables the generation of valuable insights within an acceptable timeframe. In the developed framework, response was improved by optimization of subgraphs with enablement of PICO elements. Özcan et.al., [130] concluded that semantic enrichment enables deep reasoning capability; in the automated framework it was achieved it by assimilation of information from generic and specific ontologies. Most of the existing work such as authors in [36] [131] tried to generate KGs using single iteration of annotation of concepts. The developed framework uses semantically enriched concepts as final one and creates KGs, and thus increased it semantically by many folds and improved its

searchability aspects and ability to provide specific and precise information in response to related queries. The robustness of KGs is further enhanced by inclusion of both taxonomic and non-taxonomic relationships and scaling the KG both horizontally and vertically.

Finally, it was demonstrated that the use of domain specific and language models for high performance feature extraction of underlying text data improves the performance in ML models, specifically Neuro-Symbolic clustering, PICO classification and RE model. The integration of deep learning and symbolic reasoning techniques has shown to improve the clustering performance significantly, particularly in domains such as biomedical research. BioBERT embedded layer and LSTM model improves the performance PICO classification task by 11 % in terms of accuracy for the both covid-19 dataset and cerebral aneurysm dataset. The RE model performance also improves significantly when BioBERT is used in conjunction with Bi-LSTM and CNN.

6.1 Conclusion

The use of KGs in the healthcare domain has shown to be a successful method for mapping relationships between data that are extremely diverse and structured. In addition, it provides uncanny capability to model latent relationships between information sources and capture linked information which other data models fail to do. The framework for automated KG curation was designed, incorporating a pipeline to gather data from a large corpus of research papers. Various methods were employed, including clustering of documents and its optimization using Neuro-Symbolic methods, concept extraction with the assistance of the BioPortal API, classification of PICO elements and additional Aim category using RNNs and BioBERT encoding, extraction of

taxonomic relationships using ontologies, and extraction of non-taxonomic relationships using a deep learning architecture consisting of Bi-LSTM and CNN with BioBERT representation layer. The framework presented here offers researchers the flexibility to build upon and expand its capabilities, as well as utilize specific components for various applications like food preparation, recommendation systems, and the fisheries industry. Additionally, a dedicated dataset of PICOs (Patient/Problem, Intervention, Comparison, and Outcomes) was developed specifically for the COVID-19 domain.

In the developed framework, the variations of baseline models and embeddings are used to test its different components. The concept extraction using BioPortal API has acceptable precision and recall, the clustering approach is optimized using the elbow method utilizing SSE metric, PICO classification using LSTM and BioBERT embedding shows 11 % improvement in accuracy over the baseline models on both the dataset. The use of large language model embedding in conjunction with LSTM models and external knowledge infusion in Neuro-Symbolic clustering outperforms the traditional k-means clustering algorithm and has 40% more precision in clustering documents. The current relationship extraction model using CNN, Bi-LSTM, and BioBERT encoding shows improvement in recall by 4 % on other baseline models.

6.2 Future Research

The architecture of the knowledge graph framework presented here is tested on two healthcare datasets and is flexible enough to be applied to other domains. In future, a key area of focus will be the parallelization of the data processing pipeline across multiple stages of data curation to enhance the efficiency and scalability of framework by using distributed data engineering framework such as Spark. Additionally, the

prioritization lies in the development of a user interface with interactive visualization capabilities. This user interface will empower users to explore the data through drill-down scenarios, enhancing their understanding of the data and its relationships. Furthermore, efforts will be dedicated to extending the application of the architecture framework across various domains, including but not limited to the food supply chain, dietary recommendations, agriculture, and fisheries. By exploring these diverse areas, the aim is to leverage the potential benefits of the approach and address specific challenges unique to each domain. This expansion will broaden the applicability and impact of our framework, contributing to advancements in multiple industries. As of now, the RE model for non-taxonomic relationship has the limitation of predicting only 125 different types of relationship, and in future with more data and manual annotation of relationship, this could be increased significantly. In addition, the concept extraction module is evaluated on a randomized sample of 100 abstracts due to the effort needed in manual annotation, so adding more abstracts could improve the evaluation. A further improvement to the KG completion process could be made through embedding and link prediction. The Neuro-Symbolic clustering model is evaluated on randomized sample of 100 research articles. To overcome the need for large amounts of labeled data, the LSTM network could be further developed using few-shot learning experiments.

APPENDIX

SUPPORTING PUBLICATIONS

Publications supporting proposed research:

[1] Malik, K. M., Krishnamurthy, M., Alobaidi, M., Hussain, M., Alam, F., & Malik, G. (2020). Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Systems with Applications*, *145*, 113120.

[2] Afzal, M., Alam, F., Malik, K. M., & Malik, G. M. (2020). Clinical context–aware biomedical text summarization using deep neural network: model development and validation. *Journal of medical Internet research*, *22*(10), e19810.

[3] Alam, F., Afzal, M., & Malik, K. M. (2020, January). Comparative analysis of semantic similarity techniques for medical text. In *2020 International Conference on Information Networking (ICOIN)* (pp. 106-109). IEEE.

[4] Malik, K. M., Krishnamurthy, M., Alam, F., Zakaria, H., & Malik, G. M. (2021). Introducing the Rupture Criticality Index to compare risk factor combinations associated with aneurysmal rupture. *World neurosurgery*, *146*, e38-e47.

[5] Malik, K., Alam, F., Santamaria, J., Krishnamurthy, M., & Malik, G. (2022). Toward Grading Subarachnoid Hemorrhage Risk Prediction: A Machine Learning-based Aneurysm Rupture Score. *World Neurosurgery*.

[6] Alam, F., Giglou, H. B., & Malik, K. M. Automated Clinical Knowledge Graph Generation Framework for Evidence Based Medicine. *Available at SSRN 4276561*.

[7] Alam, F., Ananbeh, O., Malik, K. M., Al Odayani, A., Hussain, I. B., Kaabia, N., & Al Aidaroos, A. (2023). Towards Predicting Length of Stay and Identification of Cohort Risk Factors Using Self-Attention Based Transformers and Association Mining: Covid-19 as Phenotype.

[8] Alam, F., Malik, K. M., & Krishnamurthy, M (2023) NeuroClustr-Empowering Biomedical Text Clustering with Neuro-Symbolic. *Intelligence International Joint Conference on Artificial Intelligence 2023 Workshop on Knowledge-Based Compositional Generalization*

Table A. Publication relevant to current research

	Research Articles	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Research Component	Domain Knowledge	Y			Y	Y		Y	
	Knowledge Extraction	Y			Y	Y	Y	Y	Y
	Knowledge Discovery		Y				Y		Y
Problem Addressed	Automated Framework			Y	Y	Y	Y	Y	Y
	New information ingestion						Y		Y
	Evidence Based Medicine			Y			Y		Y
	PICO Classification		Y						Y

REFERENCES

[1] Kang, T., Turfah, A., Kim, J., Perotte, A., & Weng, C. (2021). A neuro-symbolic method for understanding free-text medical evidence. Journal of the American Medical Informatics Association, 28(8), 1703-1711.

[2] Drancé, M., Boudin, M., Mougin, F., & Diallo, G. (2021). Neuro-symbolic XAI for Computational Drug Repurposing. In KEOD (pp. 220-225).

[3] Lenat, D. (1995). A large-scale investment in knowledge infrastructure. Communications of the ACM, 33-38.

[4] Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 211-226

[5] Oram, P. (2001). WordNet: An electronic lexical database. Christiane Fellbaum (Ed.). *Applied Psycholinguistics*, 131-134.

[6] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. The semantic web, 722-735.

[7] Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. *16th international conference on World Wide Web*, (pp. 697-706).

[8] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 78-85.

[9] Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., & Xiao, Y. (2017). CN-DBpedia: A never-ending Chinese knowledge extraction system. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, (pp. 428-438).

[10] Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. *ACM SIGMOD international conference on management of data*, (pp. 481-492).

[11] “Geonames.” GeoNames, www.geonames.org/. Accessed 19 June 2023.

[12] Du, J., & Li, X. (2020). A knowledge graph of combined drug therapies using semantic predications from biomedical literature: Algorithm development. *JMIR medical informatics,*.

[13] Wang, L., Xie, H., Han, W., Yang, X., Shi, L., Dong, J., & Wu, H. (2020). Construction of a knowledge graph for diabetes complications from expert-reviewed clinical evidences. *Computer Assisted Surgery*=, 29-35.

[14] Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C., & Wang, F. (2020). Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal*, 2737-2750.

[15] Mohamed, S. K., Nováček, V., & Nounu, A. (2020). Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 603-610.

[16] Li, N., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2020). KGHC: a knowledge graph for hepatocellular carcinoma. *BMC Medical Informatics and Decision Making*, 1-11.

[17] Johnson, J. M., Narock, T., Singh-Mohudpur, J., Fils, D., Clarke, K. C., Saksena, S., & Yeghiazarian, L. (2022). Knowledge graphs to support real-time flood impact evaluation. *AI Magazine*, 40-45.

[18] Wang, H., Wu, Z., Jiang, H., Huang, Y., Wang, J., Kopru, S., & Xie, T. (2021). Groot: An event-graph-based approach for root cause analysis in industrial settings. *36th IEEE/ACM International Conference on Automated Software Engineering (ASE)* , (pp. 419-429).

[19] Chen, C., Cui, J., Liu, G., Wu, J., & Wang, L. (2020). Survey and open problems in privacy preserving knowledge graph: Merging, query, representation, completion and applications. *arXiv preprint arXiv*.

[20] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT :a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 1234-1240.

[21] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 1-9.

[22] Shen, I., Zhang, L., Lian, J., Wu, C. H., Fierro, M. G., Argyriou, A., & Wu, T. (2020). In search for a cure: recommendation with knowledge graph on CORD-19. *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* , (pp. 3519-3520).

[23] Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., & Mungall, C. J. (2021). KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns*.

[24] Xi, J., Ye, L., Huang, Q., & Li, X. (2021). Tolerating data missing in breast cancer diagnosis from clinical ultrasound reports via knowledge graph inference. *27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, (pp. 3756-3764).

[25] Dai, Y., Guo, C., Guo, W., & Eickhoff, C. (2021). Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Briefings in bioinformatics*.

[26] Malik, K. M., Krishnamurthy, M., Alobaidi, M., Hussain, M., Alam, F., & Malik, G. (2020). Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Systems with Applications,*, 145.

[27] Yu, Y., Huang, K., Zhang, C., Glass, L. M., Sun, J., & Xiao, C. (2021). SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics,*, 2988-2995.

[28] "UC Library Guides: Evidence-Based Practice in Health: Introduction." Introduction - Evidence-Based Practice in Health - UC Library Guides at University of Canberra, canberra.libguides.com/c.php?g=599346. Accessed 19 June 2023.

[29] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[30] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[31] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[32] Chai, X. (2020). Diagnosis method of thyroid disease combining knowledge graph and deep learning. *IEEE Access*, 149787-149795.

[33] Pham, T. T. (2022). Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-Based Systems*.

[34] Dai, Y., Guo, C., Guo, W., & Eickhoff, C. (2021). Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Briefings in bioinformatics*.

[35] Mohammadhassanzadeh, H., Abidi, S. R., Van Woensel, W., & Abidi, S. S. (2018). Investigating plausible reasoning over knowledge graphs for semantics-based health data analytics. *27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* , (pp. 148-153).

[36] Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports*, 1-11.

[37] Liu, W., Yin, L., Wang, C., Liu, F., & Ni, Z. (2021). Multitask healthcare management recommendation system leveraging knowledge graph. *Journal of Healthcare Engineering*.

[38] Yu, T., Li, J., Yu, Q., Tian, Y., Shun, X., Xu, L., & Gao, H. (2017). Knowledge graph for TCM health preservation: Design, construction, and applications. *Artificial intelligence in medicine*, 48-52.

[39] Postiglione, M. (2021). Towards an Italian Healthcare Knowledge Graph. *International Conference on Similarity Search and Applications* , (pp. 387-394).

[40] Gyrard, A., Gaur, M., Shekarpour, S., Thirunarayan, K., & Sheth, A. (n.d.). Personalized health knowledge graph. *CEUR workshop proceedings* .

[41] Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., & Liu, Y. (n.d.). Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine*.

[42] Mohit, B. (2014). Named entity recognition. . *Natural language processing of semitic languages* , 221-245.

[43] Kuhn, P., Mischkewitz, S., Ring, N., & Windheuser, F. (2016). Type inference on Wikipedia list pages. *Informatik* .

[44] Smirnova, A., & Cudré-Mauroux, P. (2018). Relation extraction using distant supervision. *: A survey. ACM Computing Surveys (CSUR),* .

[45] Al-Moslmi, T., Ocaña, M. G., Opdahl, A. L., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 32862-32881.

[46] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems, ,* 494-514.

[47] Chen, Y., Argentinis, J. E., & Weber, G. (2016). IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics*, 688-701.

[48] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv*.

[49] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstration*, (pp. 55-60).

[50] Gatta, R., Vallati, M., Lenkowicz, J., Rojas, E., Damiani, A., Sacchi, L., & Valentini, V. (2017). Generating and comparing knowledge graphs of medical processes using pMineR. *Knowledge Capture Conference* , (pp. 1-4).

[51] Rastogi, N., & Zaki, M. J. (2020). Personal Health Knowledge Graphs for Patients. *arXiv preprint arXiv*.

[52] Huang, Z., Yang, J., Harmelen, F. V., & Hu, Q. (2017). Constructing knowledge graphs of depression. *International conference on health information science* , (pp. 149-161).

[53] Ma, Y., Tresp, V., & Daxberger, E. A. (2019). Embedding models for episodic knowledge graphs. *Journal of Web Semantics*.

[54] Zhang, Y., Sheng, M., Zhou, R., Wang, Y., Han, G., Zhang, H., & Dong, J. (2020). HKGB: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Information Processing & Management*.

[55] Chen, I. Y., Agrawal, M., Horng, S., & Sontag, D. (2020). Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledge graph. *In PACIFIC SYMPOSIUM ON BIOCOMPUTING*, 19-30.

[56] Wise, C., Ioannidis, V. N., Calvo, M. R., Song, X., Price, G., Kulkarni, N., & Karypis, G. (2020). COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv*.

[57] Chen, C., Ebeid, I. A., Bu, Y., & Ding, Y. (2020). Coronavirus knowledge graph: A case study. *arXiv preprint arXiv*.

[58] Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., & Onyshkevych, B. (2020). COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv*.

[59] Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction: A survey. *arXiv preprint arXiv*.

[60] Aydar, M., Bozal, O., & Ozbay, F. (2020). Neural relation extraction: a survey. . arXiv e-prints, arXiv.

[61] Daelemans, W., & Bosch, A. (2005). Memory-Based Language Processing . *Studies in Natural Language Processing*.

[62] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (pp. 1003-1011).

[63] Santus, E., Biemann, C., & Chersoni, E. (2018). Combining vector-, pattern-and graph-based information to identify discriminative attributes. *arXiv preprint arXiv:1804.11251*.

[64] Brychcín, T., Hercig, T., Steinberger, J., & Konkol, M. (2018). Uwb at semeval-2018 task 10: Capturing discriminative attributes from word distributions. *Proceedings of The 12th International Workshop on Semantic Evaluation* , (pp. 935-939).

[65] Lai, S., Leung, K. S., & Leung, Y. (2018). SUNNYNLP at SemEval-2018 Task 10: A support-vector-machine-based method for detecting semantic difference using taxonomy and word embedding features. *Proceedings of the 12th international workshop on semantic evaluation*, (pp. 741-746).

[66] Zeng, W., Lin, Y., Liu, Z., & Sun, M. (2016). Incorporating relation paths in neural relation extraction.

[67] Speer, R., & Lowry-Duda, J. (2018). Luminoso at semeval-2018 task 10: Distinguishing attributes using text corpora and relational knowledge. *arXiv preprint arXiv:1806.01733*.

[68] Buscaldi, D., Schumann, A. K., Qasemizadeh, B., Zargayouna, H., & Charnois, T. (2017). Semantic relation extraction and classification in scientific papers. *International Workshop on Semantic Evaluation*, 679-688.

[69] Liu, C., Sun, W., Chao, W., & Che, W. (2013). Convolution neural network for relation extraction. *International conference on advanced data mining and applications* , (pp. 231-242).

[70] Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. *Proceedings of the 2015 conference on empirical methods in natural language processing*, (pp. 1753-1762).

[71] Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural relation extraction with selective attention over instances. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.*

[72] Ji, G., Liu, K., He, S., & Zhao, J. (2017). Distant supervision for relation extraction with sentence-level attention and entity descriptions. *Proceedings of the AAAI conference on artificial intelligence* .

[73] Jat, S., Khandelwal, S., & Talukdar, P. (2018). Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv*.

[74] Du, J., Han, J., Way, A., & Wan, D. (2018). Multi-level structured self-attentions for distantly supervised relation extraction. *arXiv preprint arXiv:1809.00699.*

[75] Zhang, D., & Wang, D. (2015). Relation classification via recurrent neural network. . *arXiv preprint arXiv:1508.01006.*

[76] Wang, L., Cao, Z., De Melo, G., & Liu, Z. (2016). Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006.*

[77] Lee, J., Seo, S., & Choi, Y. S. (2019). Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. . *Symmetry*.

[78] Xiao, M., & Liu, C. (2016). Semantic relation classification via hierarchical recurrent neural network with attention. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* , (pp. 1254-1263).

[79] 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*.

[80] Wu, S., & He, Y. (2019). Enriching pre-trained language model with entity information for relation classification. *Proceedings of the 28th ACM international conference on information and knowledge management* , (pp. 2361-2364).

[81] Eberts, M., & Ulges, A. (2021). An end-to-end model for entity-level relation extraction using multi-instance learning. *arXiv preprint arXiv:2102.05980.*

[82] Kim, T., Yun, Y., & Kim, N. (2021). Deep learning-based knowledge graph generation for COVID-19. Sustainability, 13(4), 2276.

[83] Anastasiu, D. C., Tagarelli, A., & Karypis, G. (2018). Document clustering: the next frontier. In Data Clustering (pp. 305-338). Chapman and Hall/CRC.

[84] Fard, M. M., Thonet, T., & Gaussier, E. (2020). Seed-guided deep document clustering. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42 (pp. 3-16). Springer International Publishing.

[85] Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. Briefings in Bioinformatics, 22(1), 393-415.

[86] Davagdorj, K., Park, K. H., Amarbayasgalan, T., Munkhdalai, L., Wang, L., Li, M., & Ryu, K. H. (2022, January). Biobert based efficient clustering framework for biomedical document analysis. In Genetic and Evolutionary Computing: Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computing, October 21-23, 2021, Jilin, China (pp. 179-188). Singapore: Springer Nature Singapore.

[87] GABRALLA, L., & CHIROMA, H. (2020). Deep learning for document clustering: a survey, taxonomy and research trend. Journal of Theoretical and Applied Information Technology, 98(22), 3602-3634.

[88] Sheth, A., Gaur, M., Kursuncu, U., & Wickramarachchi, R. (2019). Shades of knowledge-infused learning for enhancing deep learning. *IEEE Internet Computing*, *23*(6), 54-63.

[89] Gaur, M., Gunaratna, K., Bhatt, S., & Sheth, A. (2022). Knowledge-Infused Learning: A Sweet Spot in Neuro-Symbolic AI. IEEE Internet Computing, 26(4), 5-11.

[90] Aspis, Y., Broda, K., Lobo, J., & Russo, A. (2022, July). Embed2Sym-Scalable Neuro-Symbolic Reasoning via Clustered Embeddings. In Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (Vol. 19, No. 1, pp. 421-431).

[91] Venugopal, D., Rus, V., & Shakya, A. (2021, June). Neuro-Symbolic Models: A Scalable, Explainable Framework for Strategy Discovery from Big Edu-Data. In Proceedings of the 2nd Learner Data Institute Workshop in Conjunction with The 14th International Educational Data Mining Conference.

[92] Kursuncu, U., Gaur, M., & Sheth, A. (2019). Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning. arXiv preprint arXiv:1912.00512.

[93] Bui, D. D. A., Del Fiol, G., Hurdle, J. F., & Jonnalagadda, S. (2016). Extractive text summarization system to aid data extraction from full text in systematic review development. Journal of biomedical informatics, 64, 265-272.

[94] Boudin, F., Shi, L., & Nie, J. Y. (2010). Improving medical information retrieval with pico element detection. In Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings 32 (pp. 50-61). Springer Berlin Heidelberg.

[95] Huang, K. C., Chiang, I. J., Xiao, F., Liao, C. C., Liu, C. C. H., & Wong, J. M. (2013). PICO element detection in medical text without metadata: Are first sentences enough?. Journal of biomedical informatics, 46(5), 940-946.

[96] Jin, D., & Szolovits, P. (2018, July). Pico element detection in medical text via long short-term memory neural networks. In Proceedings of the BioNLP 2018 workshop (pp. 67-75).

[97] Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011, December). Automatic classification of sentences to support evidence based medicine. In BMC bioinformatics (Vol. 12, No. 2, pp. 1-10). BioMed Central.

[98] Demner-Fushman, D., & Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics, 33(1), 63-103.

[99] Zlabinger, M., Andersson, L., Hanbury, A., Andersson, M., Quasnik, V., & Brassey, J. (2018, May). Medical entity corpus with PICO elements and sentiment analysis. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[100] Chabou, S., & Iglewski, M. (2015, June). PICO Extraction by combining the robustness of machine-learning methods with the rule-based methods. In 2015 World Congress on Information Technology and Computer Applications (WCITCA) (pp. 1-4). Ieee.

[101] Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M., & Marshall, I. J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. The Journal of Machine Learning Research, 17(1), 4572-4596.

[102] Boudin, F., Nie, J. Y., Bartlett, J. C., Grad, R., Pluye, P., & Dawes, M. (2010). Combining classifiers for robust PICO element detection. BMC medical informatics and decision making, 10(1), 1-6.

[103] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. ArXiv.

[104] Smileslab. “EBM_Automated_KG/Covid_pico_dataset at Main · Smileslab/EBM_AUTOMATED_KG.” GitHub, github.com/smileslab/EBM_Automated_KG/tree/main/Covid_PICO_Dataset. Accessed 19 June 2023.

[105] Xing, R., Luo, J., & Song, T. (2020). BioRel: towards large-scale biomedical relation extraction. BMC bioinformatics, 21, 1-13.

[106] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.

[107] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

[108] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., ... & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37(suppl_2), W170-W173.

[109] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.

[110] Keras, keras.io/preprocessing/text/. Accessed 19 June 2023.

[111] “Who Covid-19 Rapid Version CRF Semantic Data Model: NCBO Bioportal.” WHO COVID-19 Rapid Version CRF Semantic Data Model | NCBO BioPortal, bioportal.bioontology.org/ontologies/COVIDCRFRAPID. Accessed 19 June 2023.

[112] “Covid-19 Ontology: NCBO Bioportal.” COVID-19 Ontology | NCBO BioPortal, bioportal.bioontology.org/ontologies/COVID-19. Accessed 19 June 2023.

[113] “An Ontology for Collection and Analysis of COviD-19 Data: NCBO Bioportal.” An Ontology for Collection and Analysis of COviD-19 Data | NCBO BioPortal, bioportal.bioontology.org/ontologies/CODO. Accessed 19 June 2023.

[114] “COVID-19 Surveillance Ontology: NCBO Bioportal.” COVID-19 Surveillance Ontology | NCBO BioPortal, bioportal.bioontology.org/ontologies/COVID19. Accessed 19 June 2023.

[115] “The COVID-19 Infectious Disease Ontology: NCBO Bioportal.” The COVID-19 Infectious Disease Ontology | NCBO BioPortal, bioportal.bioontology.org/ontologies/IDO-COVID-19. Accessed 19 June 2023.

[116] “SNOMED CT: NCBO Bioportal.” SNOMED CT | NCBO BioPortal, bioportal.bioontology.org/ontologies/SNOMEDCT. Accessed 19 June 2023.

[117] “Medical Dictionary for Regulatory Activities Terminology (Meddra): NCBO Bioportal.” Medical Dictionary for Regulatory Activities Terminology (MedDRA) | NCBO BioPortal, bioportal.bioontology.org/ontologies/MEDDRA. Accessed 19 June 2023.

[118] “Neuro Behavior Ontology: Ncbo Bioportal.” Neuro Behavior Ontology | NCBO BioPortal, bioportal.bioontology.org/ontologies/NBO. Accessed 19 June 2023.

[119] “Neuroscience Information Framework (NIF) Standard Ontology: NCBO Bioportal.” Neuroscience Information Framework (NIF) Standard Ontology | NCBO BioPortal, bioportal.bioontology.org/ontologies/NIFSTD. Accessed 19 June 2023.

[120] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781..

[121] Jeffrey Pennington, Richard Socher and Christopher Manning, "Glove: Global vectors for word representation", Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.

[122] Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. Advances in neural information processing systems, 26.

[123] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

[124] Smileslab. “EBM_Automated_KG/Relationship_extraction at Main · Smileslab/EBM_AUTOMATED_KG.” GitHub, github.com/smileslab/EBM_Automated_KG/tree/main/Relationship_Extraction. Accessed 19 June 2023.

[125] [Xian et.al., 2017] Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4582-4591).

[126] [Abdi et al., 2010] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2, no. 4 (2010): 433-459.

[123] [Kodinariya et al., 2013] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1, no. 6 (2013): 90-95.

[128] Dumais, S. T. (2004). Latent semantic analysis. Annu. Rev. Inf. Sci. Technol., 38(1), 188-230.

[129] Su, Y., Sun, H., Sadler, B., Srivatsa, M., Gür, I., Yan, Z., & Yan, X. (2016, November). On generating characteristic-rich question sets for qa evaluation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 562-572).

[130] Özcan, F., Lei, C., Quamar, A., & Efthymiou, V. (2021, June). Semantic enrichment of data for AI applications. In Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning (pp. 1-7).

[131] Kamdar, M. R., Dowling, W., Carroll, M., Fitzgerald, C., Pal, S., Ross, S., ... & Samarasinghe, M. (2021). A Healthcare Knowledge Graph-based Approach to Enable Focused Clinical Search. In ISWC (Posters/Demos/Industry).