Structuring the knowledge for systematic information retrieval - knowledge graph and machine learning approach

By

M.F.Sajidh Ahamed

219179M

Department of Computer Science and Engineering,

University of Moratuwa, Sri Lanka

June 2025

Structuring the knowledge for systematic information retrieval - knowledge graph and machine learning approach

By

M.F.Sajidh Ahamed

219179M

This dissertation submitted in partial fulfilment of the requirements for the Degree of

MSc in Computer Science specialising in Data Science

Department of Computer Science and Engineering,

University of Moratuwa, Sri Lanka

June 2025

Abstract

The COVID-19 pandemic has led to the publication of a massive amount of research papers, making it hard for researchers to find relevant information quickly. This study aims to solve this problem by using knowledge graphs to organize and analyze data from the Kaggle CORD-19 dataset and AWS metadata. Over 401,270 PDF and 315,742 PMC JSON files were processed, supported by millions of metadata connections. Knowledge graphs were created to show relationships between topics, countries, institutions, authors, concepts, and sentiment scores, allowing researchers to explore the data in multiple ways.

A BERT-based sentiment analysis model was used to assign sentiment scores to papers, adding 32,299 new connections to the graph. These scores grouped papers based on similar tones and emotions, helped to uncover hidden patterns and trends. By integrating these insights into a combined knowledge graph, researchers can now traverse connections across metadata properties such as authors, institutions, topics, or sentiment scores, broadening the scope of discovery within the CORD-19 dataset.

Visualizations showed how papers are connected to different metadata properties, such as the countries where research originated, the institutions involved, and overlapping research themes. Concept graphs included confidence scores to show how strongly a paper is linked to a concept. Sentiment graphs added new layers of connections that go beyond traditional metadata. Statistics highlight the size and complexity of these graphs, with 453,633 country edges, 476,865 institutional edges, and 1,783,589 concept edges. Also, average connectivity per node increases after adding sentiment score to the knowledge graph.

This study shows that knowledge graphs are a powerful way to organize and explore large collections of research papers. Adding sentiment analysis improves the depth of analysis, making it easier to find valuable information and uncover new insights. This method can be applied to other fields in the future, providing a strong tool for solving global challenges by organizing and analyzing large datasets.

Declaration

I declare that this is my own work, and this dissertation does not incorporate without acknowledgment any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other media. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: _____ Date: 26/06/2025

Name: M.F.Sajidh Ahamed

The supervisor/s should certify the thesis/dissertation with the following declaration.

The above candidate has researched the master's thesis Dissertation under my supervision

Signature of the supervisor: _____ Date: 30/06/2025

Name: Dr Thanuja Ambegoda

Acknowledgment

I would like to express profound gratitude to my advisor, Dr Thanuja Ambegoda, for his invaluable support by providing guidance on selecting the areas of study and scoping it for this research study.

Further, I would like to thank all my colleagues for their help in finding relevant research material, sharing knowledge and experience, and for their encouragement. I am as ever, especially indebted to my wife and family for their love and support throughout my life. Finally, I wish to express my gratitude to all my colleagues, for the support given me to manage my MSc research work.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

The amount of data we have in this era is enormous. Hence structuring the data and retrieving the information from it is a challenging task. Situations like the COVID 19 pandemic escalates the data flow even more. Thus, turning that data into knowledge is vital, especially when there is large and diversified data. Hidden patterns and relationships can be brought to attention by structuring the knowledge of each data point and the relationship between them. Machines can be trained on a particular subject and with that knowledge it could augment our own capabilities by serving as assistive subject matter experts [1] [2].

Extensive research was carried out all over the world after the outbreak of the COVID 19. At the beginning of the COVID-19 pandemic, the virus had a specific form, but over time, it started to mutate. Thus, understanding the underlying molecular mechanism and identifying the mortality and severity of the disease became a challenging task. Also increased investment in related research made the quantity as a barrier to grasp the information at once. For example, as of April 28, 2020, at PubMed3 there were 19,443 papers related to coronavirus; as of June 13, 2020, there were 140K+ related papers, nearly 2.7K new papers per day. The resulting knowledge bottleneck contributed significant delays in the development of vaccines and drugs for COVID-19 [3].

Interdisciplinary research on COVID-19 provides a broad spectrum of knowledge, and connecting these insights helps accelerate the process of finding solutions to the pandemic. In building this knowledge base, the AWS COVID-19 Open Research Dataset plays a key role, offering a vast collection of research papers. This includes over 9,000 machine-readable papers on COVID-19, SARS-CoV-2, and related coronaviruses, compiled by Enigma into a single JSON file. Additionally, a comprehensive metadata file is available, listing more than 44,000 research papers on coronaviruses and COVID-19. By integrating this scientific research into a unified knowledge graph, it becomes possible to support downstream applications such as machine learning tasks, hypothesis-driven searches, and interactive user interfaces that help researchers explore and uncover key relationships.

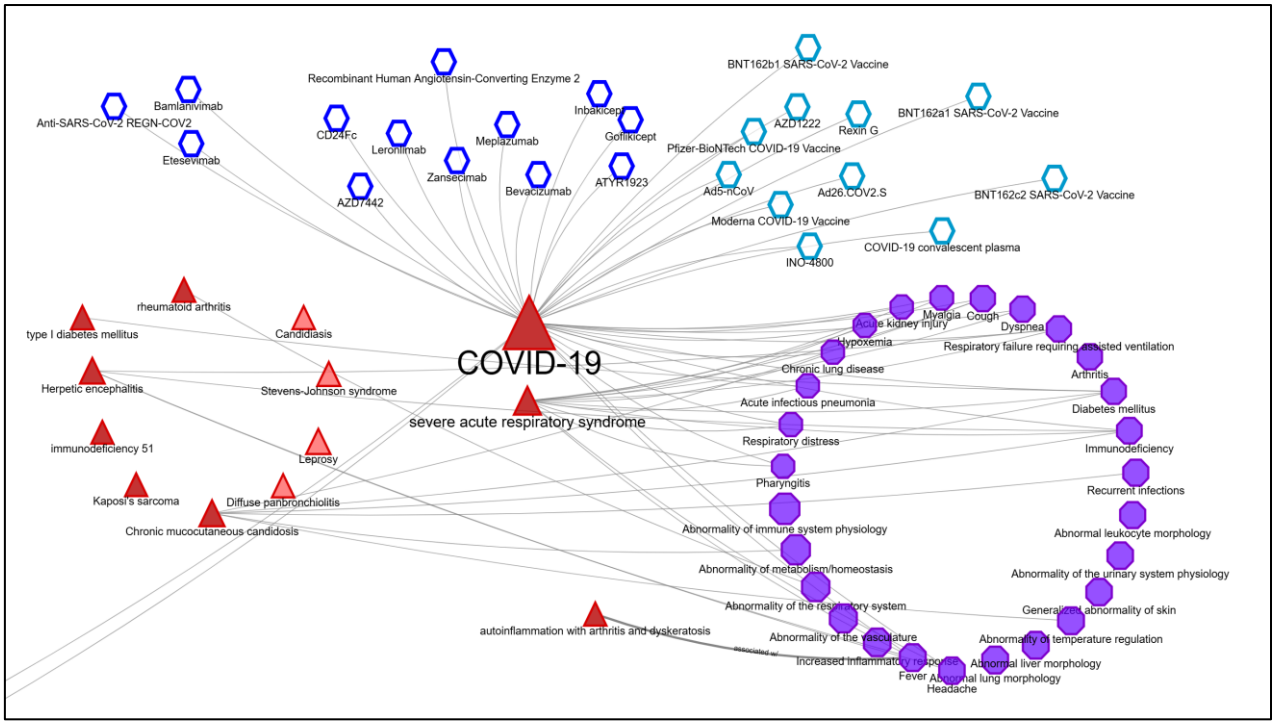


Figure 1-1 Sample Knowledge graph created using COVID -19 data (Molecular relationship) https://covid19.tubitak.gov.tr/en/covid-19-bilgi-grafikleri/,

1.1 **Research Problem**

Most existing knowledge graphs on COVID-19 primarily concentrate on medical aspects. They are typically constructed using scientific literature and serve purposes such as retrieving research articles, identifying potential drug treatments, or generating customized graphs. However, the pandemic also had a profound impact on people's daily lives, employment, and the global economy; areas that are often overlooked in these knowledge graphs.

Also, many past studies stop after building the knowledge graph or only use it in a simple way. They don't explore more advanced techniques to get deeper insights. This research aims to fill that gap by adding sentiment analysis to the knowledge graph using machine learning models like BERT. By doing this, the graph includes more useful information, like the tone or emotion in a paper. This helps group papers with similar messages or conclusions, making it easier for users to explore and understand the literature.

1.2 Research Objective

Corona virus has spread around the world in a very short span. Also, it has an ability to mutate hence researchers need to consider the findings from different geographical areas. Using knowledge graphs to organize information can improve systematic information retrieval from large datasets. This approach can also help uncover hidden patterns in COVID data more efficiently.

The novelty of this research is that it adds sentiment analysis scores to the knowledge graph. By including sentiment analysis, we provide extra context, which allows for a deeper exploration of the literature. This not only adds more information to the graph but also helps reveal hidden connections and patterns that go beyond basic metadata. The goal is to improve how knowledge graphs are used by offering a richer, sentiment-based way to explore and analyse COVID-19 literature.

- To construct a knowledge graph on CORD 19 dataset.
- Create separate knowledge graphs for different properties to identify the behaviour of the papers.
- Sentiment analysis on full paper text and assign score for each paper (adding more properties to the KG).
- Combine all the knowledge graphs for better information retrieval.

1.3 Organization of the report

This dissertation is organized into 5 chapters, including the introductory chapter. Chapter 1 provides a comprehensive introduction to the research, outlining the background, research objectives, and the significance of the study. In Chapter 2, a thorough literature review is presented, offering an in-depth exploration of the key concepts and existing research related to the study. Chapter 3 outlines the methodology used in the research, detailing the steps and techniques implemented in constructing the knowledge graphs and conducting sentiment analysis. Chapter 4 includes the analysis and the results. Finally, Chapter 5 by summarizing the key findings, elaborating the implications of the research.

Chapter 2 Literature Review

2.1 Covid-19 A global pandemic

2.1.1 Introduction and Background

The new coronavirus SARS-CoV-2, which causes COVID-19, first appeared as a respiratory disease in Wuhan, China's Hubei Province, in December 2019. The World Health Organization (WHO) proclaimed it a global pandemic in March 2020 after it quickly spread from a cluster of pneumonia patients. The virus killed around 5 million people worldwide and infected over 250 million individuals by October 2021 (WHO, 2021). In addition to placing a strain on healthcare systems, this extraordinary crisis exposed weaknesses in the world's pandemic preparedness.

One of the main causes of SARS-CoV-2's quick spread has been its transmission dynamics. In the early phases of the pandemic, studies have estimated the basic reproduction number (R0) to be between 2 and 3, suggesting the possibility of exponential growth in the absence of successful containment efforts [4]. Global travel, crowded cities, and asymptomatic transmission were some of the factors that led to the virus's widespread effects.

The emergence of SARS-CoV-2, the virus responsible for COVID-19, in Wuhan, China in December 2019 marked the beginning of a global health crisis. Declared a pandemic by the World Health Organization (WHO) in March 2020, it swiftly spread due to factors such as global mobility, urban density, and asymptomatic transmission. By October 2021, it had resulted in approximately 5 million deaths and over 250 million infections worldwide (WHO, 2021). While early estimations of the basic reproduction number (R0) ranging from 2 to 3 underscored the virus’s high transmissibility, such estimates also revealed a significant gap in real-time epidemic modeling capabilities. Although these models provided early warnings, their dependence on limited early data often led to inconsistent predictions, emphasizing the need for more dynamic and adaptable modeling frameworks.

2.1.2 Clinical Presentation and Symptoms

A wide range of clinical manifestations, from minor respiratory symptoms to severe systemic illness, are indicative of COVID-19. Fever, dry cough, exhaustion, and

occasionally anosmia (loss of smell) and dysgeusia (loss of taste) are the main symptoms. Complications include pneumonia, acute respiratory distress syndrome (ARDS), and multi-organ failure might arise from more severe instances [5]. A portion of patients, especially those with concomitant conditions like diabetes and cardiovascular disease, are at a higher risk of experiencing serious consequences, even though many recover without the need for hospitalization.

Public health officials have had difficulty identifying and isolating infected persons due to the wide range of disease presentations, including asymptomatic cases. It has been stressed that early detection and testing are essential tactics to stop the virus's spread.[6]

COVID-19 presents a spectrum of clinical manifestations, from asymptomatic to severe systemic illness. Common symptoms include fever, fatigue, and respiratory distress, with complications such as ARDS and multi-organ failure seen in high-risk groups. A strength of early clinical studies was the rapid characterization of symptoms and risk factors. However, their generalizability was often limited by geographic or demographic sampling biases. For example, many early studies lacked representation from low-income regions, leading to an incomplete global clinical picture. Furthermore, the wide variability in symptomatology, especially in asymptomatic carriers, significantly hindered efforts in disease surveillance and containment, pointing to a critical weakness in early detection protocols and the need for broader diagnostic criteria.

2.1.3 Interdisciplinary Research Efforts

Effective containment and mitigation measures are desperately needed, and this has prompted a diverse scientific response. To direct public health actions, epidemiologists have analysed the transmission dynamics of SARS-CoV-2, while virologists have investigated its structure and replication mechanisms. Research on airborne transmissibility, for example, has brought attention to the part aerosol particles play in the virus's propagation, leading to modifications in public health regulations like mask requirements and enhanced ventilation requirements. [7].

The diversity of expertise brought together during this pandemic is exemplified by collaborative efforts involving over 200 scientists from fields as varied as virology, physics,

epidemiology, and engineering. These efforts aimed to better understand the virus's airborne nature and advocate for mitigation measures that addressed these findings effectively [8].

2.1.4 Challenges in Research and Knowledge Synthesis

The vast and rapidly growing body of COVID-19 literature has posed significant challenges for researchers. The sheer volume of studies-ranging from clinical trials and epidemiological models to behavioural studies-makes it difficult to synthesize knowledge and draw actionable insights. This challenge is compounded by the varying quality of research outputs, necessitating rigorous peer review and validation processes to ensure the reliability of findings [9]

Furthermore, the interdisciplinary character of COVID-19 study has brought attention to the necessity of cross-disciplinary cooperation and efficient communication. Integrating findings from behavioural science, epidemiology, virology, and other fields requires researchers to negotiate a variety of approaches and terminology.

2.1.5 Impacts Beyond Health

Beyond its immediate health effects, COVID-19 has had profound social, economic, and psychological impacts. The pandemic has exacerbated existing inequalities, disproportionately affecting vulnerable populations and straining healthcare infrastructures in low- and middle-income countries [10]. Economically, the pandemic triggered one of the most significant recessions in recent history, with millions losing their jobs and livelihoods. Psychologically, the prolonged isolation and uncertainty associated with the pandemic have led to widespread mental health issues, including anxiety and depression.

Psychologically, mental health has suffered greatly because of the ongoing uncertainty and isolation brought on by COVID-19. Increased rates of anxiety, sadness, and substance addiction have been documented in studies, especially among those who have had financial difficulties and healthcare workers [11]. It has also been determined that children and teenagers are more susceptible, and that stress and anxiety are exacerbated by disturbances to their social connections and daily routines.

2.2 Application of Knowledge graph in general

2.2.1 Introduction to Knowledge Graphs

A knowledge graph (KG), often called a semantic network, provides an organized and connected representation of knowledge by illustrating the relationships between real-world elements. The Knowledge Graph initiative, which Google notably introduced in 2012, completely changed the way search engines provide information. The Google Knowledge Graph, which prioritized "things, not strings," improved search relevance by linking information to entities and their connections rather than just textual matches [12]. This method, which provides systematic pathways for information extraction, organization, and utilization, has now become a fundamental component in many fields.

Knowledge graphs (KGs) offer structured, semantically rich representations of entities and their relationships. Google's 2012 introduction of the Knowledge Graph was a turning point in how information was contextualized for users. While early KGs significantly improved search relevance and information retrieval, their reliance on predefined ontologies and structured data sources also revealed limitations in adaptability, especially in fast-evolving domains like COVID-19 research.

2.2.2 Challenges in Knowledge Graph Construction

One of the challenges in building a solid knowledge graph is combining inconsistent and noisy data from various sources. To guarantee clearly defined relationships between entities, the process starts with data preprocessing. Normalization, deduplication, and entity disambiguation are examples of preprocessing techniques that are essential for producing high-quality knowledge representations.

Another major problem is integrating heterogeneous data from many disciplines. For example, specialized datasets and large-scale corpora like Wikipedia and Freebase frequently contain contradictions and ambiguities that call for complex resolution techniques. To guarantee semantic consistency, methods such as entity linkage and ontology alignment are frequently employed [13].

Constructing high-quality KGs requires harmonizing noisy and heterogeneous data. Strengths of existing approaches include advanced preprocessing techniques like entity

disambiguation and ontology alignment. However, current KG systems often struggle with real-time updates and the integration of non-standardized domain-specific data. This is especially problematic in emergent fields like biomedical research during pandemics, where terminology and relationships evolve rapidly, necessitating more flexible and adaptive KG frameworks.

2.2.3 Knowledge Graphs in Structured Knowledge Representation

Because knowledge graphs can connect disparate data points, they have become the industry standard for organized information representation. News stories, financial reports, and biomedical literature are just a few examples of the vast text corpora from which they have been utilized to extract and arrange knowledge. For instance, a study that employed natural language processing (NLP) pipelines to build a financial domain knowledge graph obtained excellent precision in entity and relationship extraction [14].

KGs excel in structuring large volumes of unstructured data, aiding in knowledge discovery across domains. High accuracy has been achieved using NLP pipelines, particularly in finance and biomedicine. However, these successes are largely limited to domains with rich, high-quality datasets. In contrast, in areas with less formalized data (e.g., social sciences or public policy), KGs face challenges in establishing reliable entity relationships, often requiring substantial manual intervention.

2.2.4 Applications in Key Domains

2.2.4.1. Question Answering Systems

To create intelligent question-answering systems, knowledge graphs are essential. To traverse entity relationships and deliver precise responses to user inquiries, these systems make use of KGs. For example, domain-specific KGs are used by IBM Watson to improve its cognitive capabilities in a variety of fields, including law and healthcare [15].

KGs empower intelligent question-answering systems like IBM Watson. Their strength lies in their capacity to navigate complex relationships. However, their effectiveness diminishes with incomplete or sparsely connected graphs, limiting performance in niche or rapidly changing domains.

2.2.4.2. Recommender Systems

Knowledge graphs improve personalization in recommender systems by expressing product features and user preferences as entities with semantic links. KG-based recommendation engines are used by Amazon and Netflix to increase the relevance of recommendations by utilizing both explicit and implicit linkages [16].

KGs enhance recommendation precision through semantic reasoning. Nonetheless, balancing relevance with novelty remains a challenge, and explainability is often opaque to end users.

2.2.4.3. Information Retrieval Systems

By indexing material using semantic relationships rather than just keywords, KGs increase the effectiveness of information retrieval. KGs are used by search engines like Google to deliver contextual results that better reflect user intent.

2.2.4.4. Domain-Specific Applications

- **Medical Field**: To find disease patterns and possible remedies, knowledge graphs in the medical field help integrate clinical data, research publications, and patient records. To further drug development, for example, KGs have been utilized to connect genetic data with biomedical literature [17].
- **Finance**: To help with risk assessment and fraud detection, market entities and their interactions are modelled using financial knowledge graphs. Creating KGs from financial news to forecast market movements is one example of such application.
- **Cybersecurity**: To facilitate preventative actions, KGs assist in mapping the connections between threat actors, weaknesses, and possible attacks [18].
- **Education**: To create individualized learning pathways, KGs arrange curriculum materials and learning materials. KGs are used by systems such as Knowledge Space Theory to evaluate student competency and provide relevant learning resources [19].

2.2.5 Knowledge Graph Embeddings

When items and relationships are incorporated into continuous vector spaces, knowledge graphs can also be used as embeddings. Experiments utilizing Freebase and corpora such as Wikipedia and the New York Times have shown that joint embedding of entities and words in the same vector space can improve the accuracy of fact prediction [20]. Classification, grouping, and prediction are examples of downstream machine learning tasks that are made easier by this unified embedding technique.

Embedding techniques have successfully translated discrete entities and relations into continuous vector spaces, enabling downstream machine learning tasks. Studies using Freebase and Wikipedia have shown improved fact prediction accuracy. Still, the interpretability of these embeddings is limited, and they often lack transparency making them less suitable in high-stakes decision environments like healthcare or law.

2.2.6 Knowledge Graphs and Machine Learning

The capacity of knowledge graphs to facilitate end-to-end learning without feature engineering is one of its main advantages. KGs make complex information representation easier by integrating and harmonizing diverse data from multiple sources. Knowledge graphs, for instance, have proved crucial in the analysis of a variety of datasets during the COVID-19 epidemic, including as patient records, epidemiological data, and virology research. KGs facilitate a more comprehensive assessment of the virus's effect and dissemination by bringing different datasets together [21].

KGs are useful for machine learning models because they offer structured background information and make it possible to combine statistical learning and symbolic reasoning. Applications needing domain-specific knowledge, such automated drug discovery or legal reasoning, benefit greatly from this synergy.

2.2.7 Gaps in Existing COVID-19 Knowledge Graphs

Despite the rapid deployment of COVID-19-related knowledge graphs to facilitate information integration, decision support, and research acceleration, several notable **gaps** remain:

1. **Lack of Sentiment and Emotion Analysis**

   Most COVID-19 KGs are built on structured biomedical and epidemiological data but **neglect sentiment analysis**, which is crucial for understanding public perception, vaccine hesitancy, and behavioral dynamics. This limits their utility in informing public health messaging or social policy interventions.

2. **Temporal Reasoning Deficiencies**

   Few COVID-19 KGs support **temporal dynamics**, such as the evolution of symptoms, variant emergence, or policy changes over time. This restricts their effectiveness in trend analysis or forecasting.

3. **Limited Multilingual and Multicultural Integration**

   COVID-19 being a global phenomenon, KGs often fail to incorporate **non-English or culturally contextual data**, missing region-specific insights and leading to potential biases.

4. **Insufficient Integration with Social Media and News Streams**

   Many KGs are constructed from academic papers (e.g., CORD-19 corpus), with **minimal linkage to real-time media**, social signals, or misinformation trends, leaving them disconnected from ongoing societal discourse.

5. **Poor Interoperability with Clinical Systems**

   Despite efforts to link KGs with electronic health records (EHRs), issues such as **data privacy, standardization, and real-time syncing** continue to limit clinical integration and real-world usage.

6. **Static Structures and Manual Updates**

   Many early COVID-19 KGs were **manually curated or static**, leading to rapid obsolescence as the pandemic evolved. There is limited automation in updating or expanding the graphs using new data streams.

2.2.8 Future Directions

Knowledge graphs have more potential than their present uses. KGs are being used for real-time decision-making in autonomous systems, expanding their use in interdisciplinary research, and integrating them with graph neural networks (GNNs) for more potent reasoning skills. Future KG research should emphasize real-time adaptability, scalability, and interoperability. Integration with graph neural networks (GNNs) offers exciting potential for advanced reasoning, though this area is still in its infancy. Broader interdisciplinary collaborations, standardization efforts, and open-access initiatives will be critical for extending the utility of KGs in both scientific research and real-world applications

2.3 Application of Knowledge graph on COVID-19 literature

Natural Language Processing (NLP) approaches have been widely used by researchers to process COVID-19 datasets, especially the CORD-19 corpus. A noteworthy work created a secondary dataset known as CORD-NER (Named Entity Recognition) by applying information extraction methods to the CORD-19 corpus [22]. This method demonstrated how useful entity recognition is for structuring unstructured COVID-19 research. Given the diversity of the COVID-19 literature, figuring out how these studies relate to one another greatly helps researchers find new insights. A potent technique for structuring knowledge, knowledge graphs (KGs) have made it easier to create interconnected representations of the COVID-19 literature. These KGs have been used in many different fields, including medication repurposing, literature searches, and disease mechanism comprehension.

COVID*GRAPH, the COVID-19 Pathophysiology Knowledge Graph, and Yahoo's COVID-19 KG are notable instances of such initiatives. These projects demonstrate the development and potential of Knowledge Graphs in tackling pandemic-related issues. These graphics must be updated and modified frequently when new data becomes available due to the dynamic nature of the pandemic [23]. For example, to maintain the relevance and actionability of the knowledge graphs, the dynamic evolution of COVID-19 variations has necessitated the incorporation of genomic and proteomic data.

During evaluations with clinical professionals, a study demonstrated the creation of a knowledge base mechanism that performed better in literature searches than PubMed [24].

This illustrates how KGs can incorporate semantic links between things to enhance information retrieval. A different framework that can be tailored to different use cases was created to generate personalized graphs for the COVID-19 answer. Applications for drug repurposing, for example, may make use of protein data connected with licensed medications, whereas biomarker applications may make use of gene expression data related to certain pathways [25].

To automatically convert large corpora of scientific literature into structured, ordered, and actionable knowledge graphs, a novel system known as COVIDKG (COVID Knowledge Graph) was suggested. Researchers and physicians can select candidate investigation directions and concentrate on hypothesis testing by using COVIDKG to retrieve reliable and significant answers from scientific literature [26]. In a similar vein, another study used the CORD-19 dataset to create a COVID-19 Knowledge Graph, showcasing its value in giving academics and policymakers timely and useful information. To improve the effectiveness of information retrieval and hypothesis development, this study used machine learning-based entity extraction models to find links between medical entities.

2.3.1 Knowledge Graph Construction for COVID-19 Research

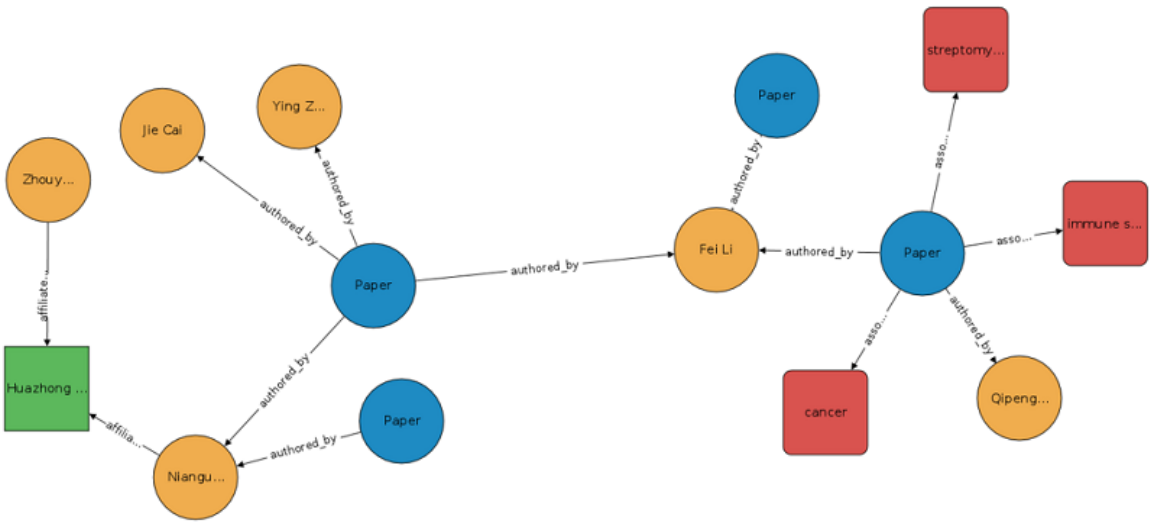


Figure 2-1 Sample Knowledge graph created using AWS data (https://aws.amazon.com/blogs/database/building-and-querying-the-aws-covid-19-knowledge-graph/)

Figure 2-1 showcases an example knowledge graph created using AWS data, which visualizes relationships between authors, papers, institutions, and extracted medical concepts. For example, there are connections between articles (blue nodes) written by scholars (yellow nodes) connected to organizations (green nodes). Amazon Comprehend Medical Detect Entities V2 is used to extract medical ideas (red nodes), which include medical conditions, drug dosages, anatomical details, treatment methods, and medications. An expansion of the Latent Dirichlet Allocation model is used to create additional topic nodes that organize documents according to their content. The relationships between different items in the COVID-19 literature are better understood thanks to these representations [27].

A knowledge graph of 10 entity types and 160 research articles was created to improve comprehension of the virus's pathogenesis. This graph included thorough coverage of biological processes, drug-target interactions, and virus-related genes and proteins. The initiative demonstrated how KGs can bring disparate studies into coherent knowledge models, which helps researchers find important discoveries more quickly [21].

2.3.3 Comparative Overview of Prominent COVID-19 Knowledge Graphs

Table 2-1 Comparative Overview of Prominent COVID-19 Knowledge Graphs

Knowledge Graph/CORD-19	**Data Sources**	**Strengths**	**Key Limitations**
CORD-19 KG [39]	CORD-19 scholarly articles	Strong biomedical focus; NLP-based relation extraction	No sentiment or social media integration; limited clinical linkage

COVID-KG [26]	PubMed	Entity linking; ontology-based construction	Static; lacks real-time updates and multilingual support
KG-COVID-19 [25]	Biomedical ontologies, drug/gene databases	High semantic rigor; useful for drug repurposing	Complex to use; no integration of behavioural data
COVID Graph [38]	Clinical, epidemiological, literature	Interdisciplinary; supports queries across domains	Poor visualization tools; limited temporal capabilities
CoVex (COVID-19 Explorer) [40]	Protein interactions, scientific publications	Interactive exploration of protein–drug–gene relationships	No socio-economic or mental health data; no user sentiment
OntoCovid [41]	Ontologies and structured datasets	Robust ontology-driven modeling	Narrow scope; lacks real-world data inputs

2.3.3 Applications Beyond Literature Search

Knowledge graphs have been used in COVID-19 research in ways that go beyond conventional book searches and drug discovery. To provide data that could be interpreted, a study showed how to use KGs for disease classification by combining them with more conventional techniques. The suggested framework outperformed earlier approaches in

terms of accuracy, demonstrating the possibility of combining KGs with machine learning methods for reliable and comprehensible disease classification [29].

Furthermore, Knowledge Graphs are beneficial when used properly in a Retrieval-Augmented Generation (RAG) pipeline. They can significantly enhance the contextual understanding and relevance of the generated responses. For instance, incorporating domain-specific data into KGs based on the user's query can provide more detailed and accurate information. With the help of machine learning techniques, these enriched knowledge bases can be categorized, hidden associations can be found, and new edges can be predicted. This could lead to the discovery of previously undiscovered links between different pieces of information, thus enhancing the quality and coherence of the generated responses in a RAG pipeline.

2.3.4 Recent Advances in Knowledge Graph Techniques

According to recent research, a knowledge graph and interactive visual analysis can be used to analyse the COVID-19 pandemic. It emphasizes how crucial it is to monitor patient data and relationships to enhance social isolation protocols. The study integrates semantic linkages with knowledge graph visualization and extracts entities and knowledge using the Conditional Random Field (CRF) algorithm. Knowledge graph data is stored in the Neo4j graph database. The usage of PageRank and Label propagation techniques to find community propagation in the network is also covered in the study. The findings demonstrate that it is possible to analyse the transmission mechanism, important nodes, and activity tracks by creating a knowledge graph of COVID-19 patient activity [30].

To increase their usefulness, further studies have looked on sophisticated KG construction methods. Temporal knowledge graphs, for example, have been created to monitor relationship changes over time, allowing researchers to examine how the epidemic has progressed and how it has changed public policy and healthcare systems. The informational value of COVID-19 KGs has been further enhanced by incorporating additional datasets, such as hospitalization rates and immunization statistics. Furthermore, link prediction and entity recognition have been enhanced by embedding-based techniques for knowledge graph node representation, allowing researchers to more accurately predict interactions and relationships inside the graph [31].

The creation of multi-modal Knowledge Graphs, which incorporate several data kinds like text, graphics, and structured data, is another area of study [31]. For example, combining textual literature and radiology imaging data in KGs may offer a thorough grasp of illness development, supporting clinical decision-making. Additionally, federated learning techniques have been put out to build distributed KGs while preserving data privacy, which is crucial for sensitive medical information [32].

To sum up, the use of knowledge graphs in COVID-19 research has shown enormous promise in terms of structuring and drawing conclusions from large and dispersed data. Future KGs can offer even more useful insights by combining machine learning, Large Language models, and multi-modal information, spurring advancements in healthcare research and pandemic control.

Chapter 3 Methodology

3.1 Knowledge graph Summary

An ordered and structured depiction of actual entities and the connections between them is called a knowledge graph. It provides a thorough method for modelling data that captures the intricate relationships found in the actual world. While the relationships between these things provide the context, associations, and meaning of how these entities interact or are related, these entities can represent a wide range of objects, concepts, events, or situations. Knowledge graphs are especially effective at revealing hidden insights, establishing connections, and enhancing comprehension across a variety of fields because of their interconnectedness.

A knowledge graph is typically stored in a graph database, a specialized system that natively handles the storage, retrieval, and management of graph-based data structures. Graph databases are excellent at immediately storing and exploring relationships, in contrast to standard relational databases, which arrange data into tables and necessitate numerous joins to connect related data points. Knowledge graphs are perfect for applications that require responsiveness and scalability because of their ability to effectively model, query, and navigate densely interrelated data in real time.

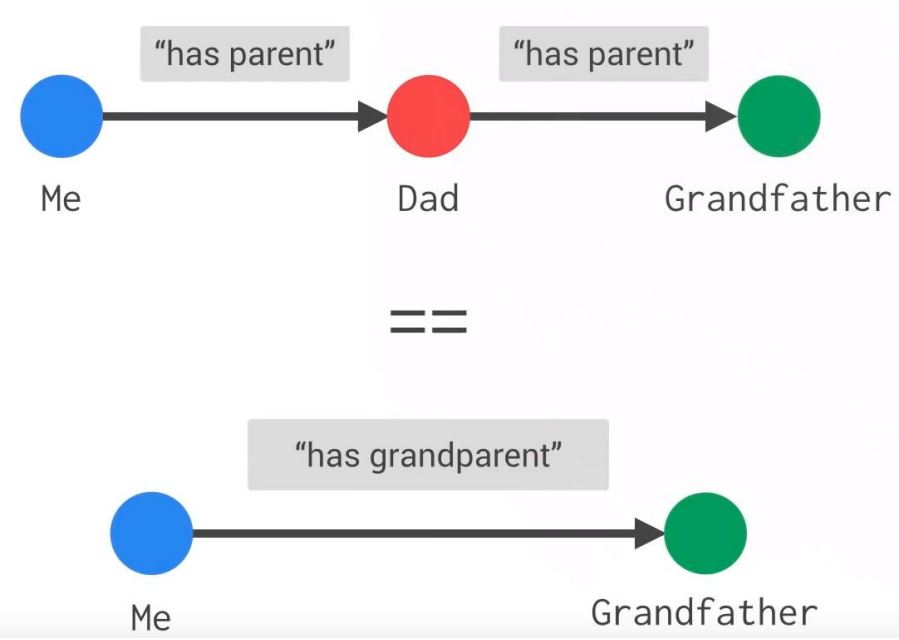


Figure 3-1 Example of a knowledge Graph equivalency

A knowledge graph's fundamental function is to integrate relationships, data, and organizing principles into a cohesive framework. These organizing principles, which provide the data a flexible and conceptual structure, might be viewed as rules, hierarchies, or ontologies. These frameworks make it possible to classify, categorize, and standardize the data consistently, guaranteeing that the graph can accurately depict intricate relationships and systems. As new information, entities, and relationships are added over time, these guidelines also facilitate the knowledge graph's growth and scalability.

By linking dissimilar data points, knowledge graphs can reveal new information and insights, which is one of its main advantages. Users can detect patterns, draw new connections, and get previously undiscovered insights by using knowledge graphs, which show not just the data but also the relationships between data points. For companies, this translates into better decision-making skills, a deeper comprehension of context, and the capacity to respond to intricate queries that transcend straightforward data retrieval.

Knowledge graphs are extremely adaptable and relevant to a wide range of usage patterns due to their architecture and adaptability. By effectively navigating relationships, knowledge graphs can provide precise and contextually relevant information in real-time applications. By offering semantic comprehension and linking searches to relevant results, they improve search engines, recommendation systems, and information retrieval in search and discovery. Furthermore, knowledge graphs are essential for providing generative AI systems with a solid foundation in the age of artificial intelligence. They enhance the precision, dependability, and contextual awareness of AI-powered question-answering systems and big language models by supplying structured and precise data.

3.2 Data Collection and Dataset Preparation

3.2.1 The CORD-19 dataset

The Allen Institute for AI, in partnership with top research teams, created the extensive COVID-19 Open Research Dataset (CORD-19). In response to the COVID-19 pandemic, this dataset was developed to help the international scientific community by giving them access to a sizable library of academic publications about the virus and the coronavirus

family. More than 29,000 of the more than 44,000 academic publications in the collection are available in full form [33].

CORD-19's main objective is to encourage scholars to use the latest developments in natural language processing (NLP) to produce fresh ideas that will help combat COVID-19. As new research is published in peer-reviewed journals and archival sites like bioRxiv and medRxiv, the dataset is updated every week [33].

Cord_uid, sha, source_x, title, doi, pmcid, pubmed_id, license, abstract, publish_time, authors, journal, microsoft_academic_paper_id, who_covidence, has_pdf_parse, has_pmc_xml_parse, full_text_file, and url are among the metadata fields that are included in the dataset's structure. Researchers may effectively access and use the information for their own study needs because to this comprehensive metadata.

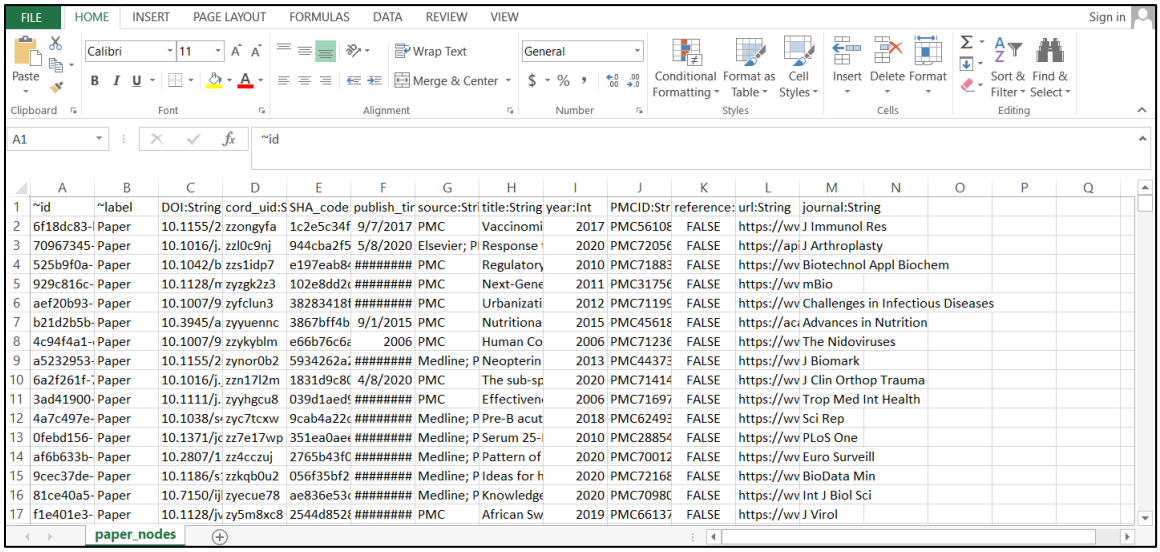
GitHub offers not only the dataset but also the source code that describes how it is collected, processed, updated, and released. Because of its transparency, researchers can better understand the dataset's processes and help make it better.

The goal of the publicly available CORD-19 dataset is to enable scholars everywhere to aid in the comprehension and mitigation of COVID-19. The Allen Institute for AI and its collaborators want to speed up scientific innovation and discovery in response to the epidemic by offering a comprehensive and constantly updated database.

3.2.2 Metadata Extraction from AWS

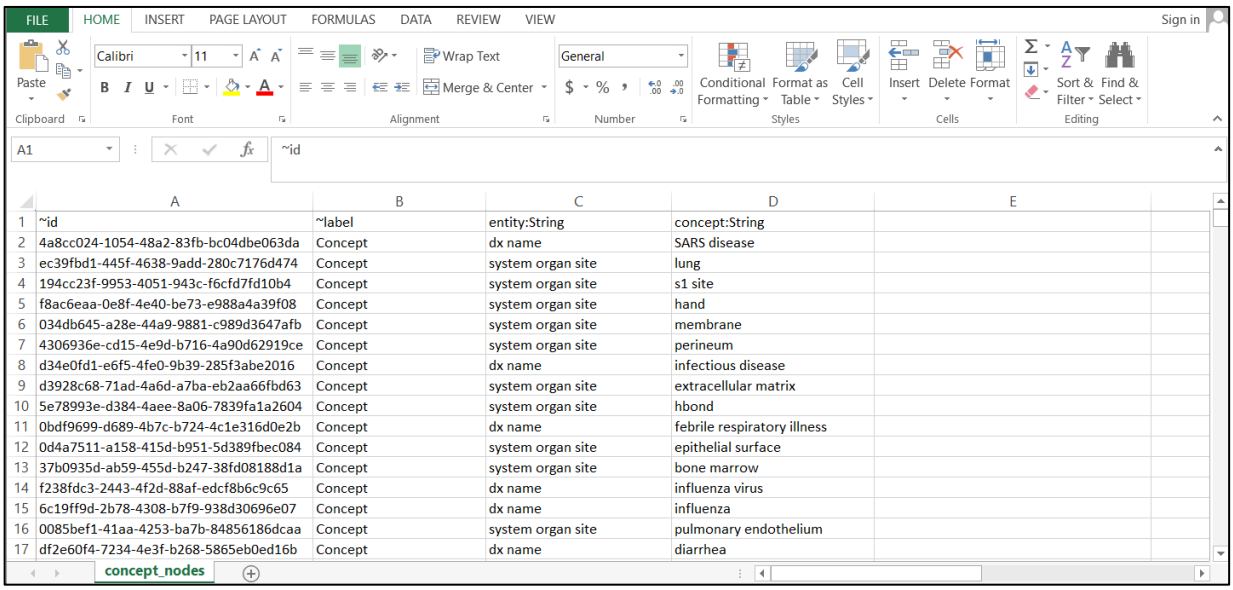
We started with downloading the data from the AWS which consists of the paper metadata. For example, the nodes and the edges are elaborated by a csv file with the names of the authors, topics, institute country etc. This data from AWS did not contain the actual papers of the CORD-19 dataset.

Each of the CSV files lists different properties of the papers. Each row of these csv files is associated with an ID which can be used to map them to other connecting nodes. Below are some of the examples of how the metadata is represented in the CSV files and how they map to one another with the IDs.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	~id	~label	DOI:String	cord_uid:S	SHA_code	publish_tir	source:Stri	title:String	year:Int	PMCID:Str	reference:	url:String	journal:String
2	6f18dc83-	Paper	10.1155/2	zzongyfa	1c2e5c34f	9/7/2017	PMC	Vaccinomi	2017	PMC56108	FALSE	https://wv	J Immunol Res
3	70967345-	Paper	10.1016/j.	zzl0c9nj	944cba2f5	5/8/2020	Elsevier; P	Response	2020	PMC72056	FALSE	https://api	J Arthroplasty
4	525b9f0a-	Paper	10.1042/b	zzs1idp7	e197eab8	########	PMC	Regulatory	2010	PMC71883	FALSE	https://wv	Biotechnol Appl Biochem
5	929c816c-	Paper	10.1128/m	zyzgk2z3	102e8dd2	########	PMC	Next-Gene	2011	PMC31756	FALSE	https://wv	mBio
6	aef20b93-	Paper	10.1007/9	zyfclun3	38283418	########	PMC	Urbanizati	2012	PMC7119	FALSE	https://wv	Challenges in Infectious Diseases
7	b21d2b5b-	Paper	10.3945/a	zyyuennc	3867bff4b	9/1/2015	PMC	Nutritiona	2015	PMC45618	FALSE	https://ac	Advances in Nutrition
8	4c94f4a1-	Paper	10.1007/9	zzykyblm	e66b76c6	2006	PMC	Human Co	2006	PMC7123	FALSE	https://wv	The Nidoviruses
9	a5232953-	Paper	10.1155/2	zynor0b2	5934262a	########	Medline; P	Neopterin	2013	PMC44373	FALSE	https://wv	J Biomark
10	6a2f261f-	Paper	10.1016/j.	zzn17l2m	1831d9c8	4/8/2020	PMC	The sub-sp	2020	PMC71414	FALSE	https://wv	J Clin Orthop Trauma
11	3ad41900-	Paper	10.1111/j.	zyyhgcu8	039d1aed	########	PMC	Effectiven	2006	PMC71697	FALSE	https://wv	Trop Med Int Health
12	4a7c497e-	Paper	10.1038/s	zyc7tcxw	9cab4a22	########	Medline; P	Pre-B acut	2018	PMC62493	FALSE	https://wv	Sci Rep
13	0febd156-	Paper	10.1371/jc	zz7e17wp	351ea0ae	########	Medline; P	Serum 25-	2010	PMC28854	FALSE	https://wv	PLoS One
14	af6b633b-	Paper	10.2807/1	zz4cczuj	2765b43f	########	Medline; P	Pattern of	2020	PMC70012	FALSE	https://wv	Euro Surveill
15	9cec37de-	Paper	10.1186/s	zzkqb0u2	056f35bf2	########	Medline; P	Ideas for h	2020	PMC72168	FALSE	https://wv	BioData Min
16	81ce40a5-	Paper	10.7150/ij	zyecue78	ae836e53	########	Medline; P	Knowledge	2020	PMC7098	FALSE	https://wv	Int J Biol Sci
17	f1e401e3-	Paper	10.1128/jv	zy5m8xc8	2544d852	########	PMC	African Sw	2019	PMC66137	FALSE	https://wv	J Virol

Figure 3-2 Example of Paper node



	A	B	C	D	E
1	~id	~label	entity:String	concept:String	
2	4a8cc024-1054-48a2-83fb-bc04dbe063da	Concept	dx name	SARS disease	
3	ec39fbd1-445f-4638-9add-280c7176d474	Concept	system organ site	lung	
4	194cc23f-9953-4051-943c-f6cfd7fd10b4	Concept	system organ site	s1 site	
5	f8ac6eaa-0e8f-4e40-be73-e988a4a39f08	Concept	system organ site	hand	
6	034db645-a28e-44a9-9881-c989d3647afb	Concept	system organ site	membrane	
7	4306936e-cd15-4e9d-b716-4a90d62919ce	Concept	system organ site	perineum	
8	d34e0fd1-e6f5-4fe0-9b39-285f3abe2016	Concept	dx name	infectious disease	
9	d3928c68-71ad-4a6d-a7ba-eb2aa66fbd63	Concept	system organ site	extracellular matrix	
10	5e78993e-d384-4aee-8a06-7839fa1a2604	Concept	system organ site	hbond	
11	0bdf9699-d689-4b7c-b724-4c1e316d0e2b	Concept	dx name	febrile respiratory illness	
12	0d4a7511-a158-415d-b951-5d389fbec084	Concept	system organ site	epithelial surface	
13	37b0935d-ab59-455d-b247-38fd08188d1a	Concept	system organ site	bone marrow	
14	f238fdc3-2443-4f2d-88af-edcf8b6c9c65	Concept	dx name	influenza virus	
15	6c19ff9d-2b78-4308-b7f9-938d30696e07	Concept	dx name	influenza	
16	0085bef1-41aa-4253-ba7b-84856186dcaa	Concept	system organ site	pulmonary endothelium	
17	df2e60f4-7234-4e3f-b268-5865eb0ed16b	Concept	dx name	diarrhea	

Figure 3-3 Sample of Concept Node.

Above figures illustrate each entity and its properties. We have similar data for all the six categories separately.

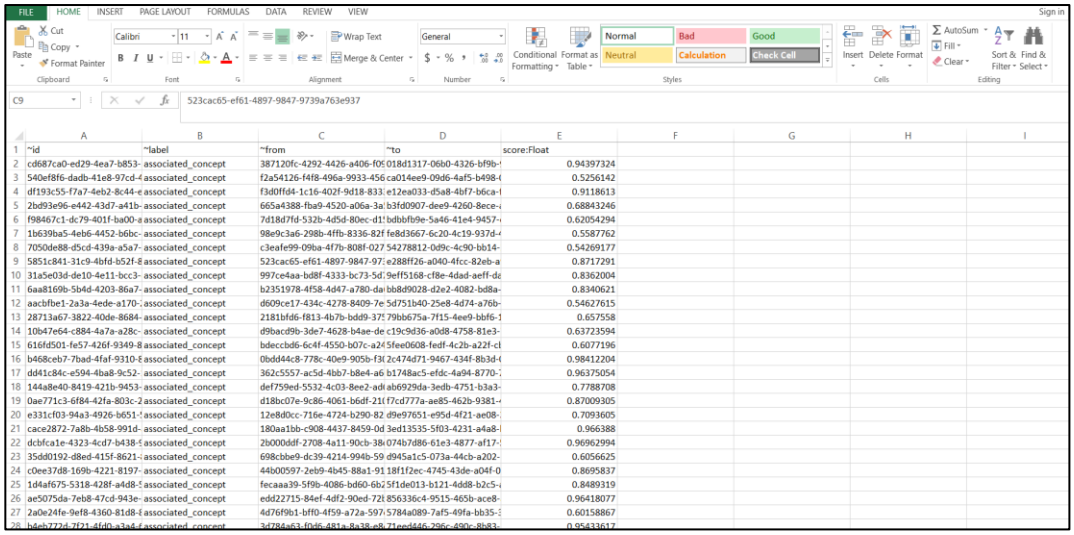


Figure 3-4 Sample relationship of paper to concept node.

This figure indicates the relationship between two entities namely paper and concept. Each paper is related to a concept and one paper can relate to one or more concepts.

For each of properties found in the metadata dataset from AWS are listed below along with their file names.

Files:

- Edges/
  - author_to_institution.csv
  - paper_to_author.csv
  - paper_to_concept.csv
  - paper_to_paper_similarity.csv
  - paper_to_reference.csv
  - Paper_to_topic.csv
- Nodes/
  - concept_nodes.csv
  - institution_nodes.csv
  - topic_nodes.csv
  - paper_author_nodes.csv
  - Paper.csv

3.2.3 Full Dataset from Kaggle

The paper version of the dataset was found in Kaggle with around 20 GB of zipped scientific papers in JSON format. Unzipping the zip file, the total amount of data was about 86 GB.

The total number of papers found in the Kaggle dataset are,

Full text:

PDF – 401,270 json

PMC (PubMed Central) – 315,742 json

Each paper was in JSON format and structured to conveniently find the reference papers, abstracts, body texts, etc.

A custom loop was implemented to extract the paper titles and their referenced paper titles from the JSON files. Iterating through the dataset required approximately two hours, resulting in a dictionary format:

```
{  "paper_title_1": ["REF01_name", "REF02_name", ...],

   "paper_title_2": ["REF01_name", "REF02_name", ...],    ...}
```

The above extracted full text of each paper was used to calculate sentiment score. Which later be converted to knowledge graph and combined with other entities.

3.3 Knowledge Graph Construction

3.3.1 Properties of the Knowledge Graph

Multiple knowledge graphs were constructed based on distinct properties derived from the metadata and extracted data. In these graphs:

- Each node represented an individual paper.
- Edges between nodes varied depending on the specific property under consideration.

The properties used to generate knowledge graphs included:

1. **Country**: Links between papers based on the country of the affiliated institution.
2. **Author**: Connections between papers sharing common authors.
3. **Topic**: Papers categorised and linked based on similar research topics.
4. **Institution**: Papers affiliated with the same research institution were connected.
5. **Concepts**: Papers sharing common scientific concepts.
6. **Sentiment Score**: Connections based on the sentiment analysis of the paper content.

For each knowledge graph, the structure remained the same, with nodes representing individual papers, while the edges represented relationships defined by the specific property in focus. This multi-faceted approach enabled a comprehensive exploration of the dataset from various perspectives.

3.3.2 Knowledge graph creation

A knowledge graph structures knowledge by representing it as interconnected nodes and edges within a graph. This approach organizes information derived from context into a network of entities and their relationships. Entities in a knowledge graph can include real-world objects, events, or abstract concepts, while relationships are depicted as arrows linking pairs of entities. Each connection, known as an edge, is typically expressed as a "triple" or a fact. This triple consists of three key components.

1. Head

2. Relation

3. Tail

When starting to construct a knowledge graph, a few related files needed to be load on to the memory for processing. For example, for the “Topics” knowledge graph, the following files were needed.

- Title_to_paper_map.json
- Topic_nodes.csv
- paper_to_topic.csv

For example:

"paper_A" → "related_to_topic" → "epidemiology"

The title to paper map, mapped the IDs from the csv files to the actual paper names in the dataset. As the other 2 files where mainly ID based, an ID to paper mapping file was needed to bridge the gap.

- Head: Paper title
- Relation: "has_topic"
- Tail: Topic name

Processing these 3 files together resulted in a Head-relation-tail file for the Topic knowledge graph. First 10 lines of the file are given below.

Table 3-1 First 10 fields of the Topics Knowledge graph file

Paper	**Score**	**Topic**
atrans-splicedleadersequenceonactinmrnaincelegans	1.0	'genomics'
chapter100angiotensin-convertingenzyme-2	0.954	'epidemiology'
technoeconomicmodelingofplant-basedgriffithsinmanufacturing	0.9927	public-health-policies
technoeconomicmodelingofplant-basedgriffithsinmanufacturing	0.8872	lab-trials-human
technoeconomicmodelingofplant-basedgriffithsinmanufacturing	0.6584	healthcare-industry
thecovid-19pandemicimplicationsforthecytologylaboratory	0.9988	public-health-policies
thecovid-19pandemicimplicationsforthecytologylaboratory	0.9399	epidemiology
extendedstorageofsars-cov2nasopharyngealswabsdoesnotnegativelyimpactresultsofmolecular-basedtesting	1.0	lab-trials-human
extendedstorageofsars-cov2nasopharyngealswabsdoesnotnegativelyimpactresultsofmolecular-basedtesting	0.9883	public-health-policies
extendedstorageofsars-cov2nasopharyngealswabsdoesnotnegativelyimpactresultsofmolecular-basedtesting	0.7607	public-health-policies

A knowledge graph which connects the papers to country was created by taking the country data from the 'institution_nodes.csv' file as a separate csv file for the country information were not available. Thus 2 knowledge graphs were created from the 'institution_nodes.csv' file.

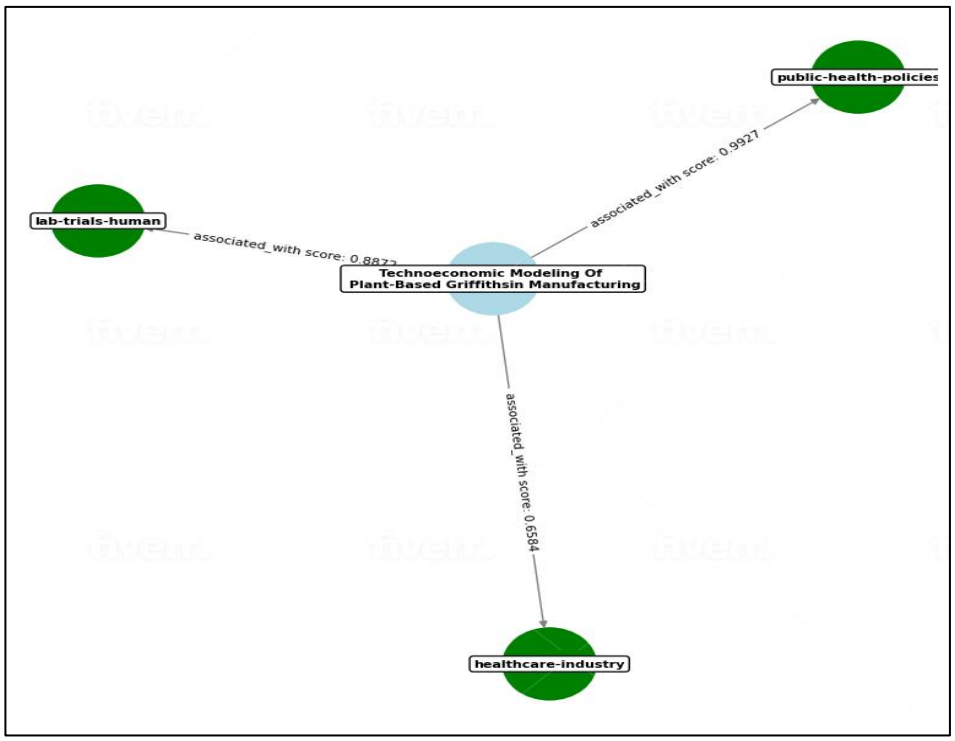


Figure 3-5 Sample of a single knowledge graph

This Figure illustrates the single knowledge graph created for a given paper with concept. Here I have picked only one paper for illustration purposes. Below we have a sample knowledge graph for multiple papers and its associated concepts. Arranging the papers in a knowledge graph helps to expedite the analysis and identify the hidden connections between papers.

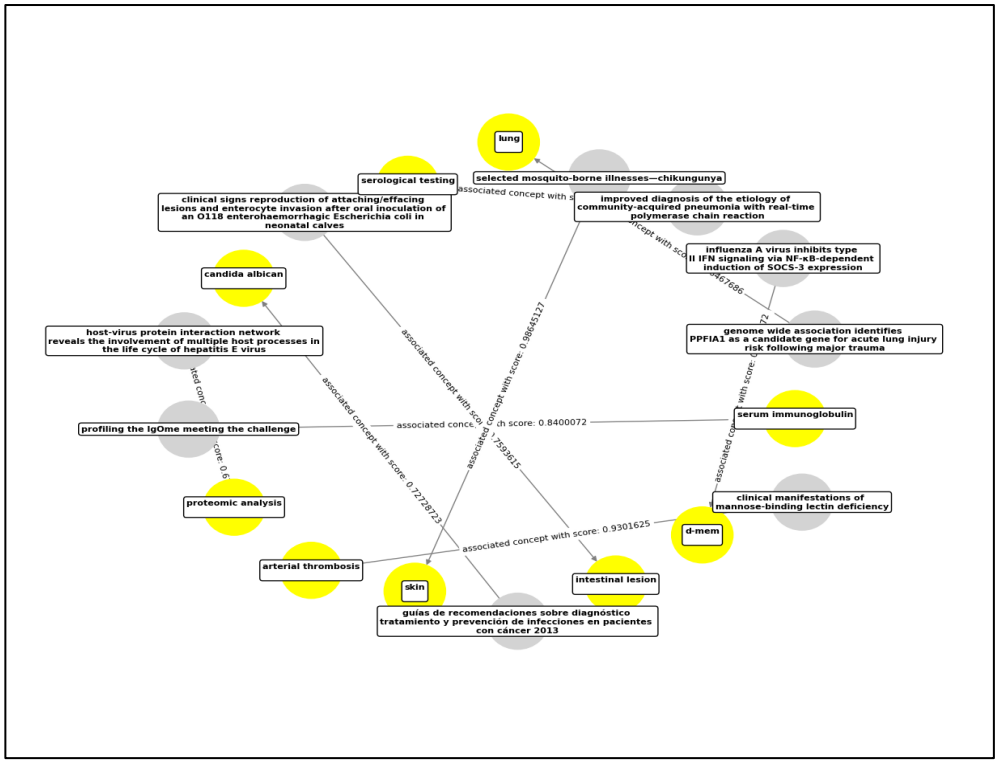


Figure 3-6 Random 10 papers with the concept it is associated with.

Combining all the properties and create a combined knowledge graph

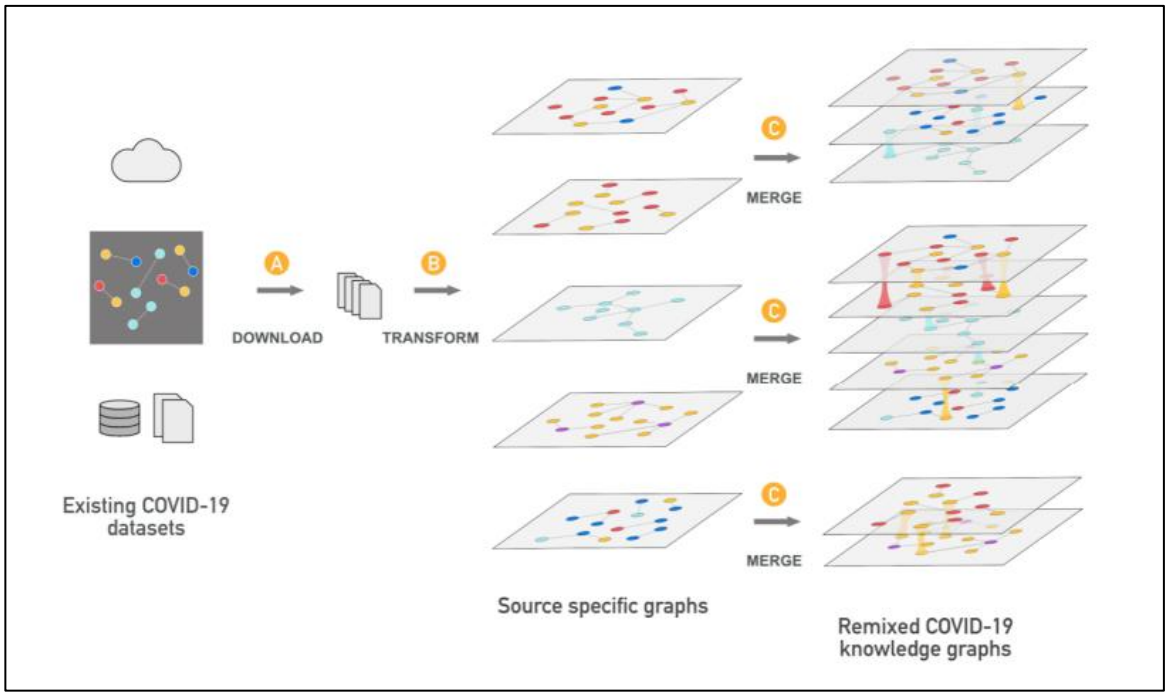


Figure 3-7 Combining all the knowledge graph into one. [25]

Above figure illustrates the method used in creating the knowledge graphs. Where initially knowledge graph was constructed on each property and then combined them together for better understanding. Since I have created separate knowledge graphs based on the title, we can combine the properties based on our need of analysis.

3.3.3 Understanding BERT and Its Role in Text Classification

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language representation model developed by **Google AI** in 2018. It revolutionized NLP by introducing a bidirectional approach to understanding text, which allowed for context from both sides of a word to be considered simultaneously [35].

3.3.3.1 Key Features of BERT

Transformer Architecture:

BERT is built on the **Transformer** architecture introduced by Vaswani et al. (2017). Transformers rely on self-attention mechanisms, which enable them to model relationships between words across long sentences.

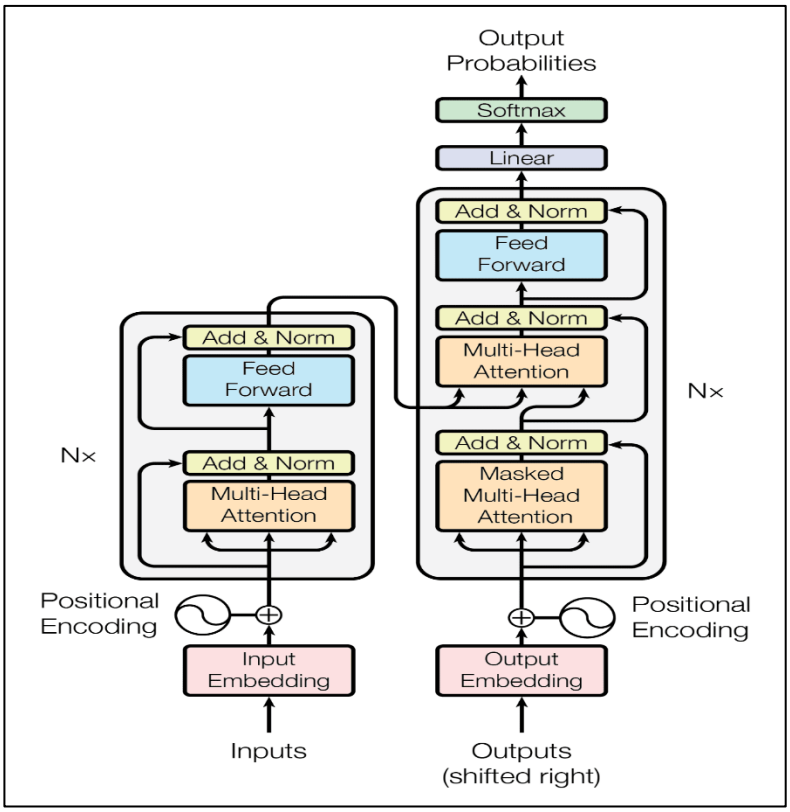


Figure 3-8 Transformer Architecture [36]

- **Self-Attention Equation:**

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices, and dk is the dimension of the keys.

Bidirectional Contextualization:

Traditional models like Word2Vec or GloVe create static embeddings, but BERT dynamically contextualizes words based on surrounding context. For instance, the word "bank" in "river bank" versus "financial bank" is correctly understood depending on context.

Pretraining Objectives:

BERT is pretrained using two objectives:

- **Masked Language Model (MLM):**
  BERT masks 15% of the input tokens and predicts them based on their context.
- **Next Sentence Prediction (NSP):**
  BERT learns relationships between sentence pairs by predicting whether the second sentence follows the first.

3.3.3.2 BERT in Text Classification

BERT excels in text classification tasks due to its ability to produce embeddings rich in semantic and syntactic information. Below is the step-by-step process:

a. Tokenization:

- BERT uses a WordPiece tokenizer that splits text into smaller subwords or tokens.
- The input sequence begins with a **[CLS]** token, representing the entire sequence for classification tasks, and ends with a **[SEP]** token.
- Example: "The cat is sleeping." → [CLS], "the", "cat", "is", "sleep", "##ing", [SEP].

b. Input Representation:

Each token is mapped to a combination of:

- **Token Embeddings:** Word-level embeddings.
- **Segment Embeddings:** To distinguish between sentence pairs.
- **Positional Embeddings:** To maintain word order.

Input is represented as:

$$Input\ Representation = E_{Token} + E_{Segment} + E_{Position}$$

c. Passing Through BERT Encoder:

- The input embeddings are processed by a series of transformer layers, each consisting of:
  - **Multi-head Self-Attention:** Captures relationships between tokens.
  - **Feedforward Neural Networks (FFN):** Adds non-linearity and learns feature transformations.

d. Classification Head:

- The final hidden state of the **[CLS]** token is passed to a feedforward layer with a softmax function to predict the class probabilities:

$$P(class \mid text) = softmax(W.h_{[CLS]} + b)$$

where $h_{[CLS]}$ is the embedding for the [CLS] token, W and b are learned parameters.

e. Fine-tuning on Classification Data:

- During fine-tuning, the model adjusts its weights on a labeled dataset specific to the classification task, such as sentiment analysis.

3.3.3.3 BERT's Strengths in Sentiment Classification

1. Deep Contextual Understanding:

BERT considers words in both forward and backward contexts, enabling nuanced sentiment detection in complex sentences [36].

2. Domain Adaptability:

With fine-tuning, BERT can adapt to domain-specific language, improving performance on specialized datasets.

3. Handling Ambiguity:

The self-attention mechanism allows BERT to resolve ambiguities in word meanings.

3.3.4 Sentiment Score

One of the key properties integrated into the knowledge graph was Sentiment Score, derived from the textual content of the papers. The JSON files contained text blocks of the abstract, body and other parts of the paper. These text blocks were extracted and divided into paragraphs. The paragraphs were then sent to a selected sentiment classfication model where it was classified into several sentiment categories with their associated confidence scores.

3.3.4.1 Sentiment Model Selection

Initial models like distilbert-base-uncased-finetuned-sst-2-english provided coarse sentiment labels (Positive/Negative). To improve granularity:

- The nlptown/bert-base-multilingual-uncased-sentiment model was used.
- It outputs five sentiment classes (1 to 5).
- Scores were scaled into a 0–100 range for finer clustering.

3.3.4.2 Sentiment Score Calculation

The sentiment score was calculated as follows:

1. Text Extraction: Text blocks from each paper were aggregated.
2. Model Inference: Each text block was processed using the sentiment model, which returned:
    - A label (e.g., 1–5), representing the sentiment class.
    - A confidence score, indicating the model's certainty for the label.
3. Score Formulation:

$$Base\ Score\ = \frac{(Label - 1)}{4} \times 100$$

$$Final\ Score\ = Base\ Score \times Confidence$$

4. Averaging: For each paper, the scores from all text blocks were averaged:

$$Average\ score\ = \sum_{n}^{i=0} \quad Final\ Score(i)\ /\ n$$

Where n is the total number of text blocks for the paper.

5. Clustering: The averaged sentiment scores were rounded and stored as whole numbers in a panda DataFrame. These values were used to generate sentiment clusters in the Sentiment Knowledge Graph.

3.3.4.3 Example Calculation

For a text block:

- Label: 4
- Confidence: 0.86

The scores were calculated as:

- Base Score = (4−1) ×4×100=75
- Final Score = 75×0.86 = 64.5

These scores were aggregated across all text blocks for each paper to compute the overall sentiment score.

The sentiment scores were taken as whole numbers to create distinct nodes for the knowledge graph. As their final score were in between 0 to 100, we are able to have 100 different nodes which each score value and connect all the papers with the same score. This would result in 100 more cluster which would be in the newly created sentiment knowledge graph.

3.3.5 Evaluation of Knowledge Graph Effectiveness

To assess the utility and quality of the knowledge graphs, the following criteria were applied:

Graph Connectivity:

Measured average degree and clustering coefficients to ensure meaningful interlinkage between nodes.

Semantic Integrity:

Spot-checked node relationships to validate correctness (e.g., whether linked papers share authors or institutions).

Coverage Metrics:

Assessed percentage of total papers represented in each graph, identifying gaps due to missing metadata.

Average connectivity per node:

How the average connectivity per node changes after sentiment score addition

Chapter 4 Analysis and Results

As stated earlier, the total number of papers which are downloaded from the Kaggle dataset are PDF – 401,270 Json and PMC (PubMed Central) – 315,742 Json files. The metadata files which were downloaded from AWS are used to create Knowledge graphs. The length of the csv files is given below.

Table 4-1 Number of lines in the available files from the AWS dataset

Files	Lines
Paper_to_topic.csv	131,509
Institution_nodes.csv	29,932
Paper_node.csv	57,313
Author_to_institution.csv	164,136
Paper_author_nodes.csv	234,827
Paper_to_author.csv	340,789
Concept_nodes.csv	76,804
Paper_to_concept.csv	1,836,969

After the knowledge graphs were constructed, the number of connections for each graph were measured. This would help us understand the complexity and massiveness of the knowledge graphs. Below are the lengths of the lines generated.

Table 4-2 Number of connections made after the formation of the knowledge graphs

Knowledge Graph	Connections
Topic	129,530
Country	476,865
Institution	476,865
Author	335,138
Concept	1,783,786
Sentiment	32,299

4.1 Visualization of Knowledge graphs

4.1.1 Topic Knowledge Graph

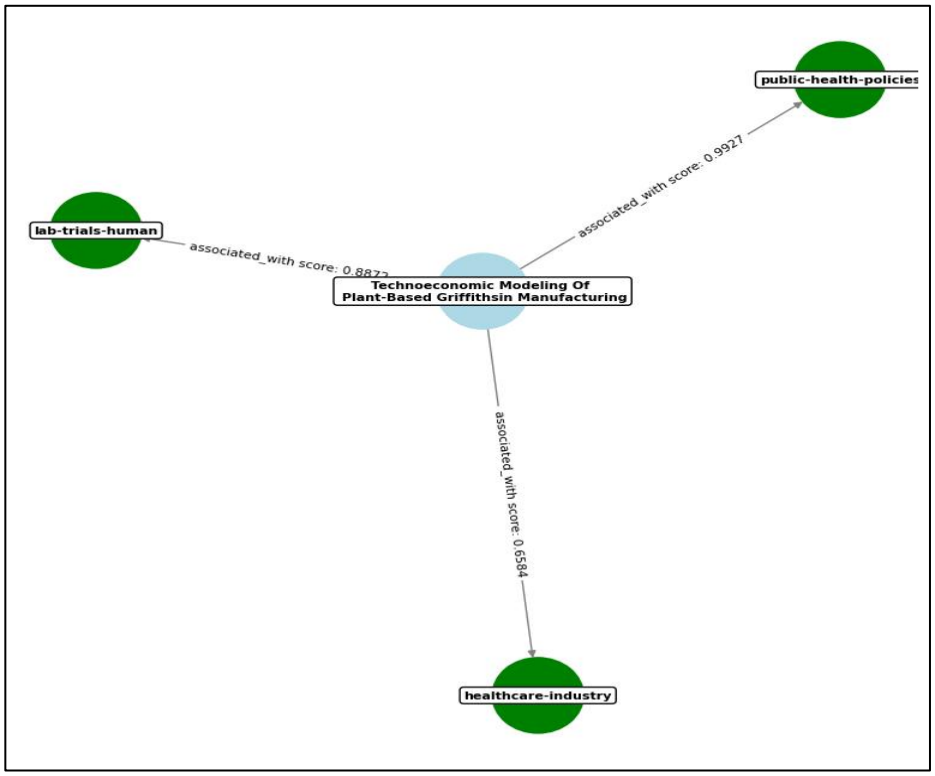


Figure 4-1 A paper connect with different topics

This visualization was done for a single paper with multiple topics which the paper talks about. And below shows how multiple papers relate to the same topics.

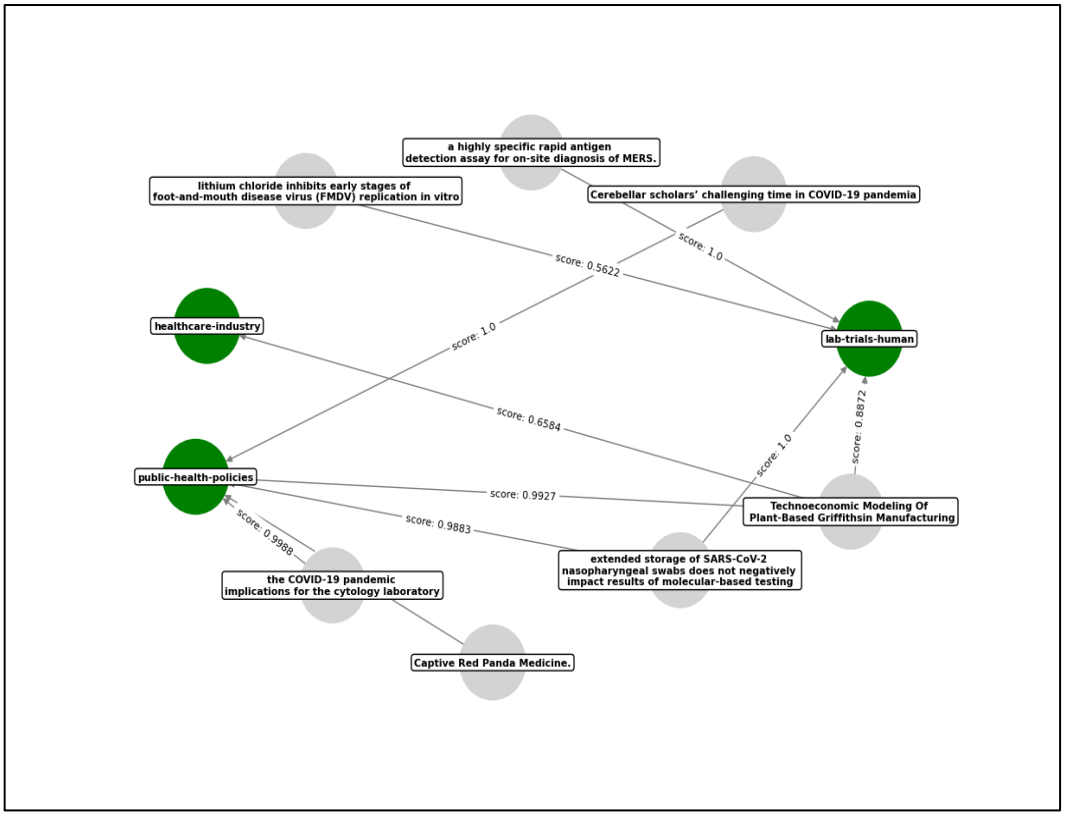


Figure 4-2 Topics which connect multiple papers together

This visualization represents a Topic Knowledge Graph where individual scientific papers (gray nodes) are connected to various topics (green nodes) such as public-health-policies, healthcare-industry, and lab-trials-human. The lines (edges) between the papers and concepts are annotated with confidence scores, which indicate how strongly each paper is associated with a given concept.

For example:

- The paper titled "the COVID-19 pandemic implications for the cytology laboratory" is linked to the public-health-policies concept with a high score of 0.9988, suggesting strong relevance.
- Another paper, "Technoeconomic Modeling of Plant-Based Griffithsin Manufacturing", is associated with lab-trials-human (score: 0.8872), public-health-policies (score: 0.9927), and healthcare-industry (score: 0.6584), reflecting its multidisciplinary nature.

This graph helps highlight how papers span multiple concepts and the varying degrees of association. By visualizing concept relationships with confidence scores, it supports more

nuanced literature exploration and enables researchers to quickly identify papers most relevant to specific conceptual themes.

4.1.2 Country Knowledge Graph

There are some papers which the publishing countries are not given. Thus, the knowledge graph for the missing country values were given a Nan.

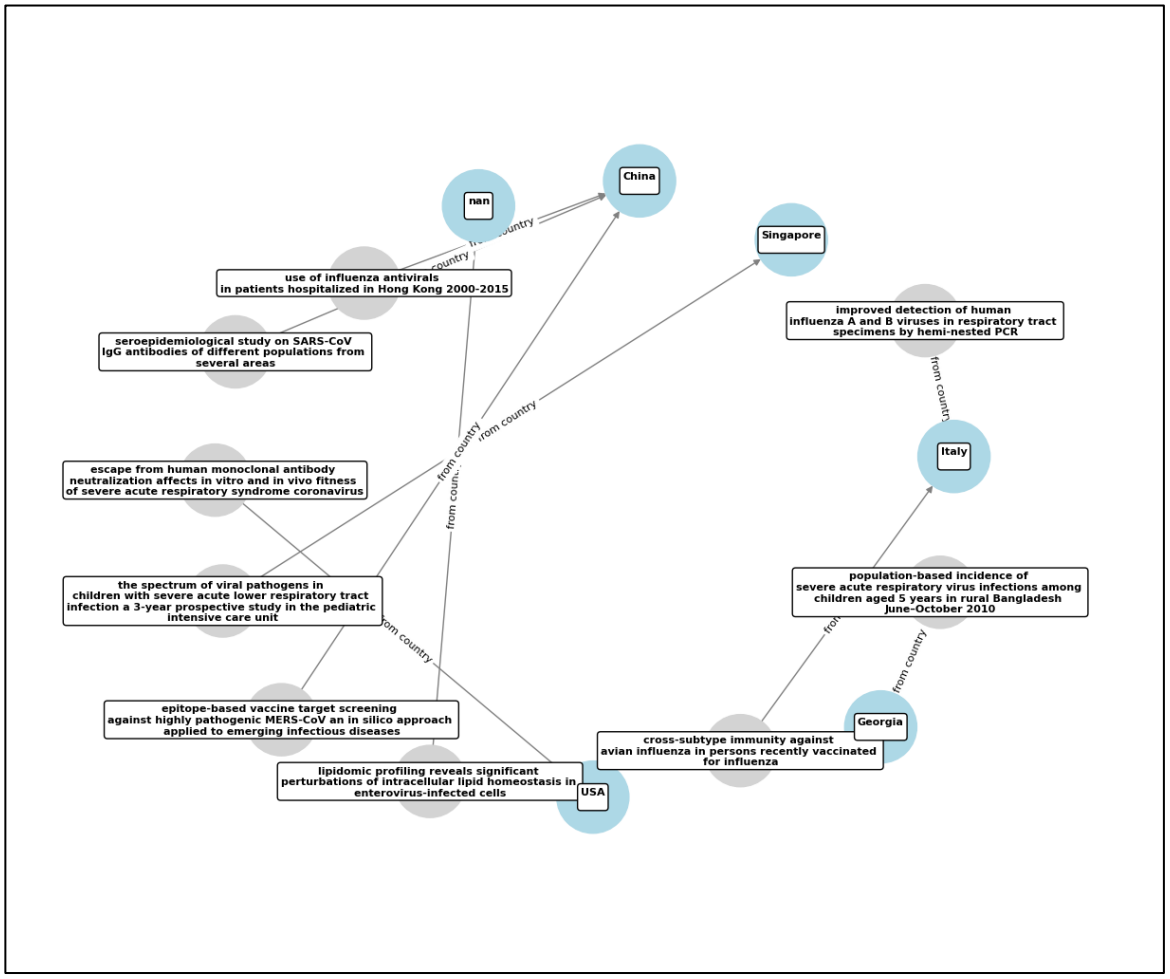


Figure 4-3 Multiple papers connection shown which seemed to originate from the same country.

4.1.3 Author Knowledge Graph

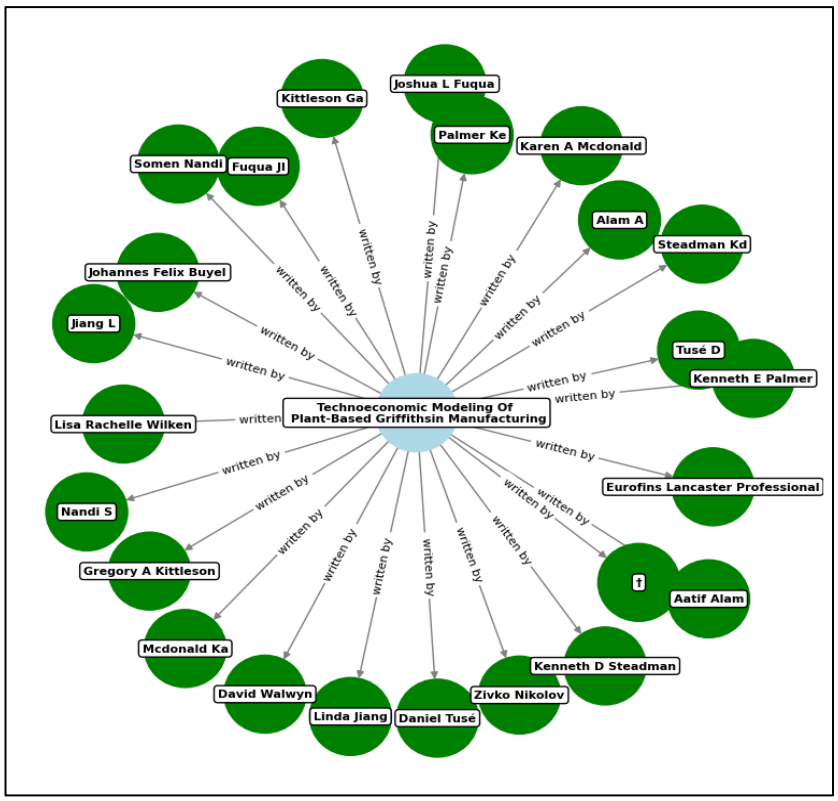


Figure 4-4 Paper shown with all its authors.

A paper normally has a lot of authors. Thus, a small part of the author connection in other paper is show below.

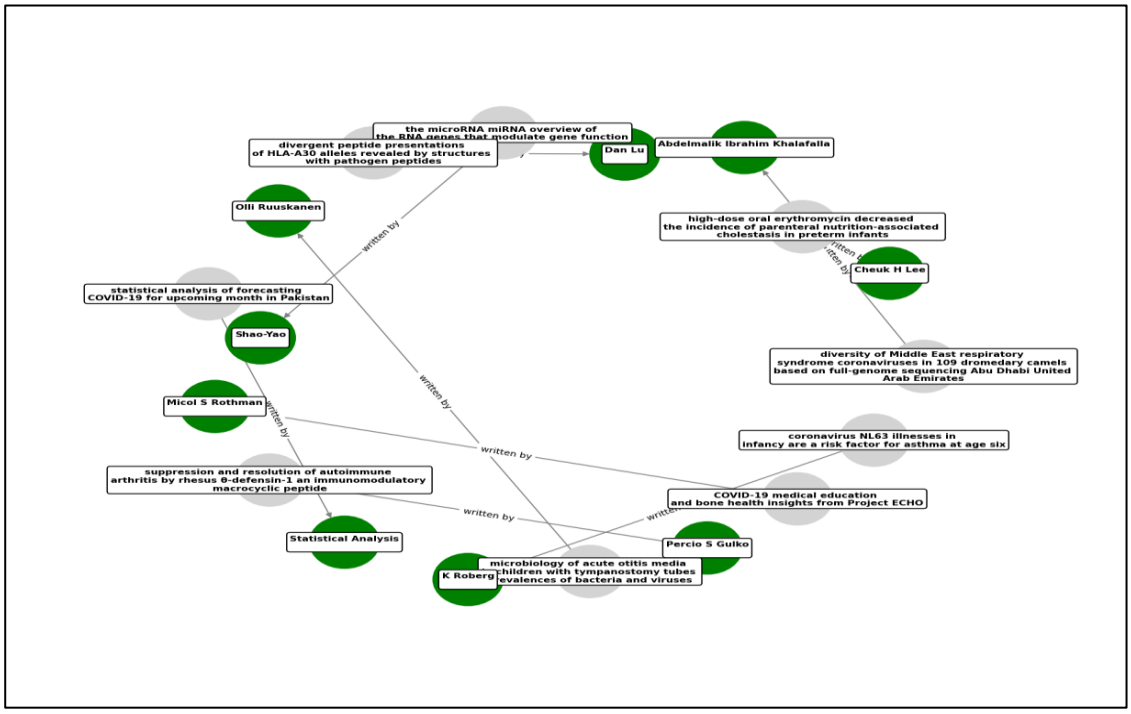


Figure 4-5 multiple papers with connected authors

4.1.4 Concept Knowledge Graph

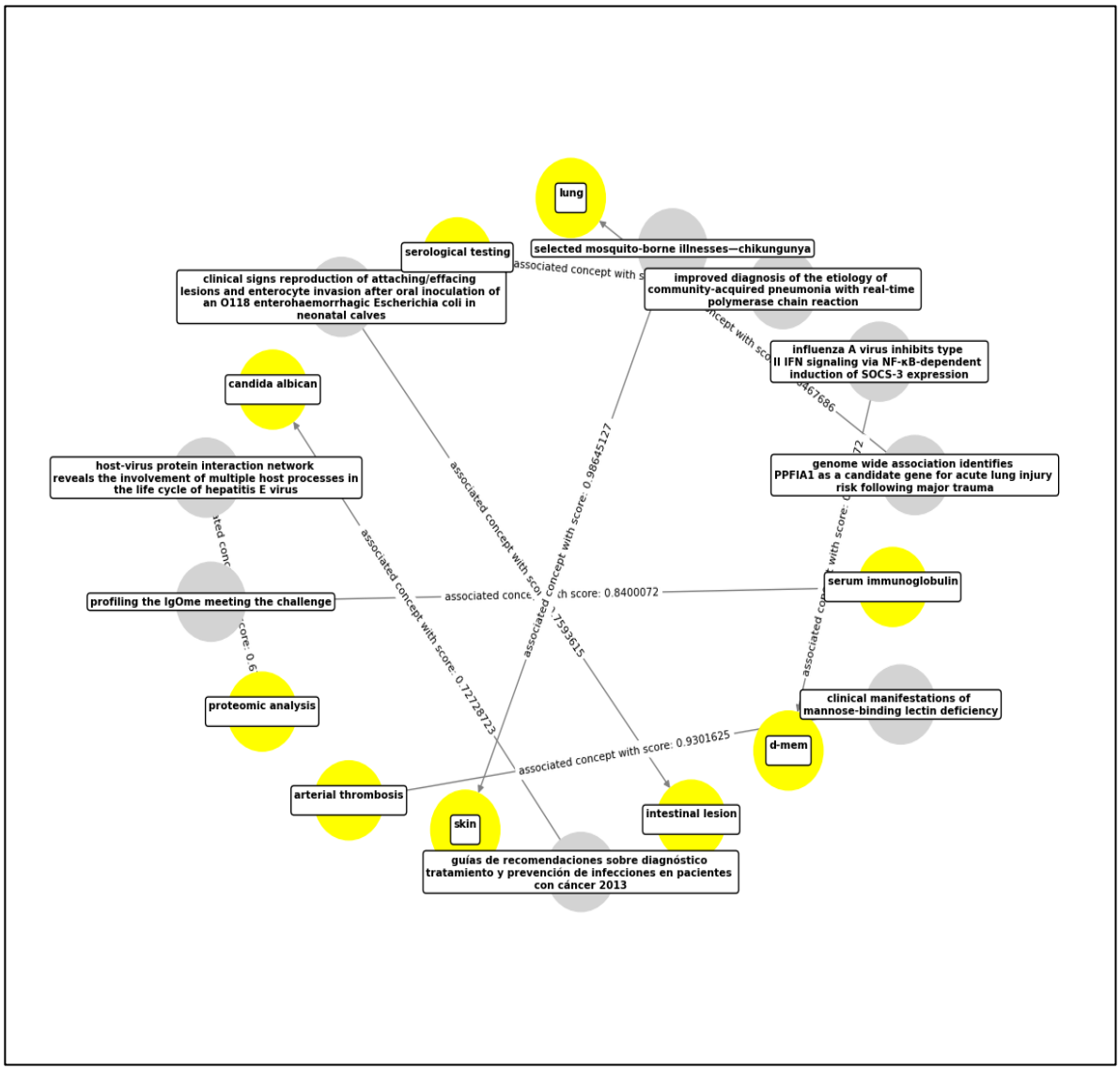


Figure 4-6 Papers connected to concepts with their associate confidence scores.

There are concept confidence scores which are added to the visualized knowledge graph for the Concepts. That is because the for the classification of the concepts in the paper, there are no sure classes but rather a score of how sure the model thinks the paper has the connect concepts.

4.1.5 Institution Knowledge Graph

Complimentary to the author graphs, the institution produces more research papers. Thus we can see that one institution is connected to multiple papers and not the other way around. Although there might be some which are produced from joint institutions but that is not found in the metadata dataset from AWS.

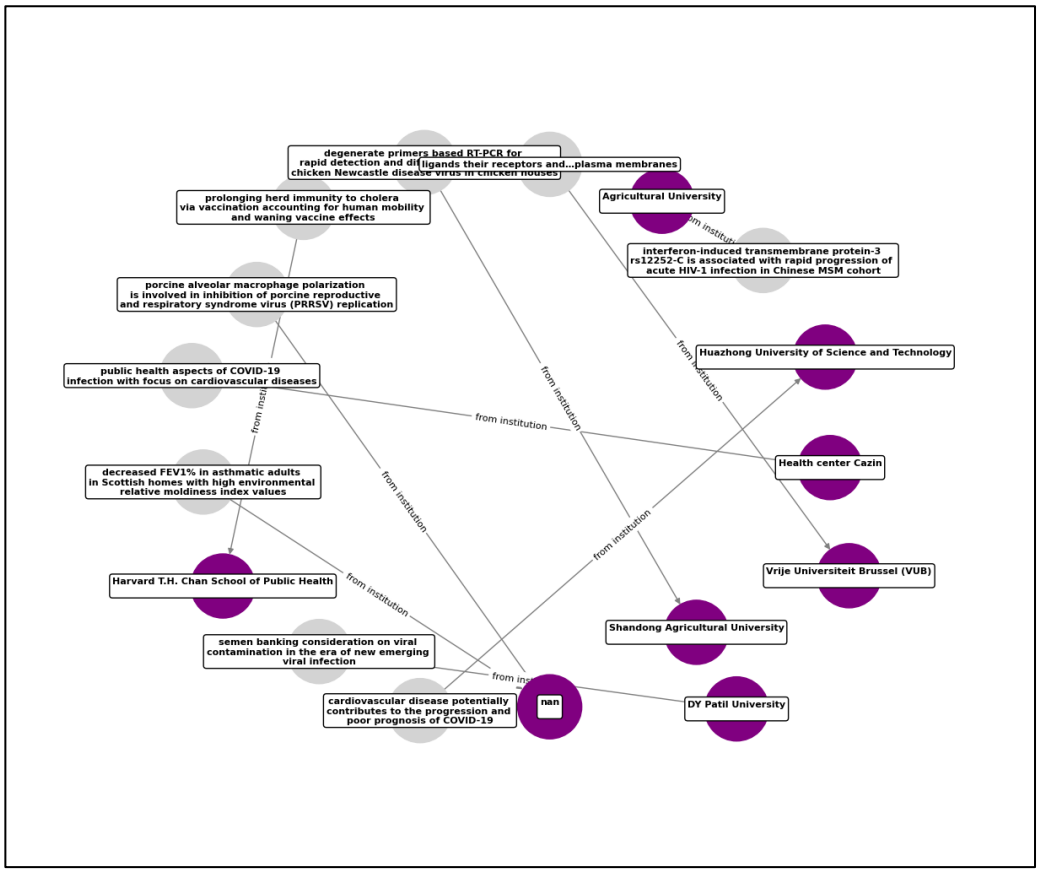


Figure 4-7 Papers connected with many Institutions.

4.1.6 Sentiment Knowledge Graph

The sentiment value was added to find new relations of the papers. Till now there are limited papers which are connected by the properties. Software services which use the existing search engine with the CORD-19 data can navigate the paper which are connected by the predefined metadata.

With the addition of the sentiment values, new connections between the papers are formed which has the same sentiment value. Thus, new papers can be discovered with the use of the sentiment value in the combined knowledge graph.

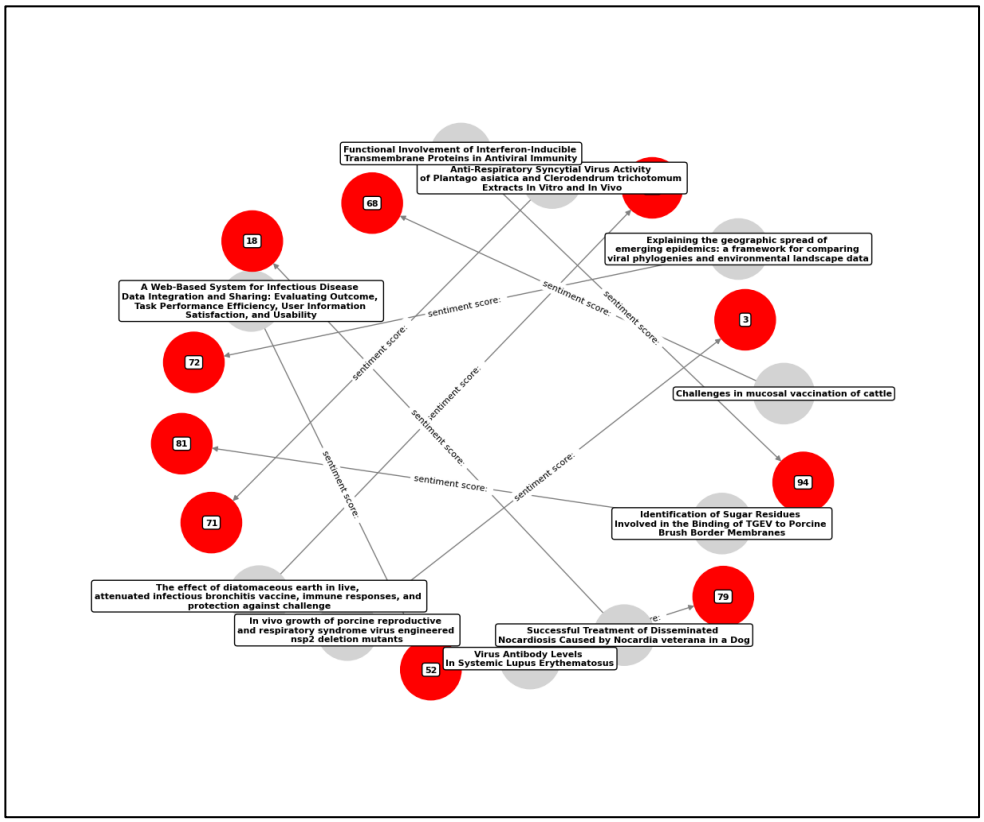


Figure 4-8 Other papers which can be discovered with the addition of the sentiment scores for all the papers.

Here we can see that many papers can be connected which has different sentiment scores. Likewise, a single score node of the sentiment knowledge graph can also connect multiple papers which shares a sentiment similar in its literature.

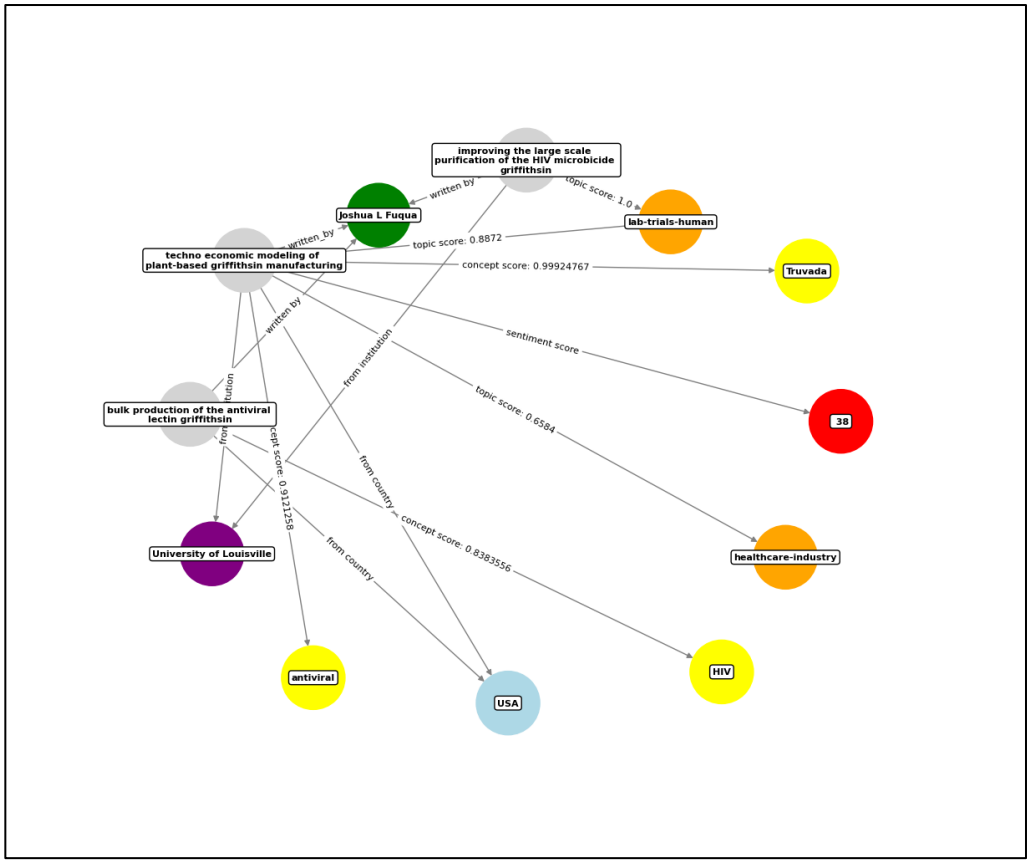


Figure 4-9 A combined knowledge graph with all the properties mentioned in this research.

Table 4-3 Color for each of the property in the combined knowledge graph

Color for each property	*Color*
Country	*Blue*
Author	*Green*
Topic	*Orange*
Institution	*Purple*
Concepts	*Yellow*
Sentiment Score	*Red*
Papers	*light Gray*

From all the above shown knowledge graphs, we can combine and form a combined knowledge graph which we are able to traverse based on their individual metadata and characterises. Thus, other papers which shares the same metadata or sentiment score are also discovered through this combine knowledge graph.

With the combined knowledge graph shown below we can see that we can get similar papers based of the metadata properties of the AWS dataset.

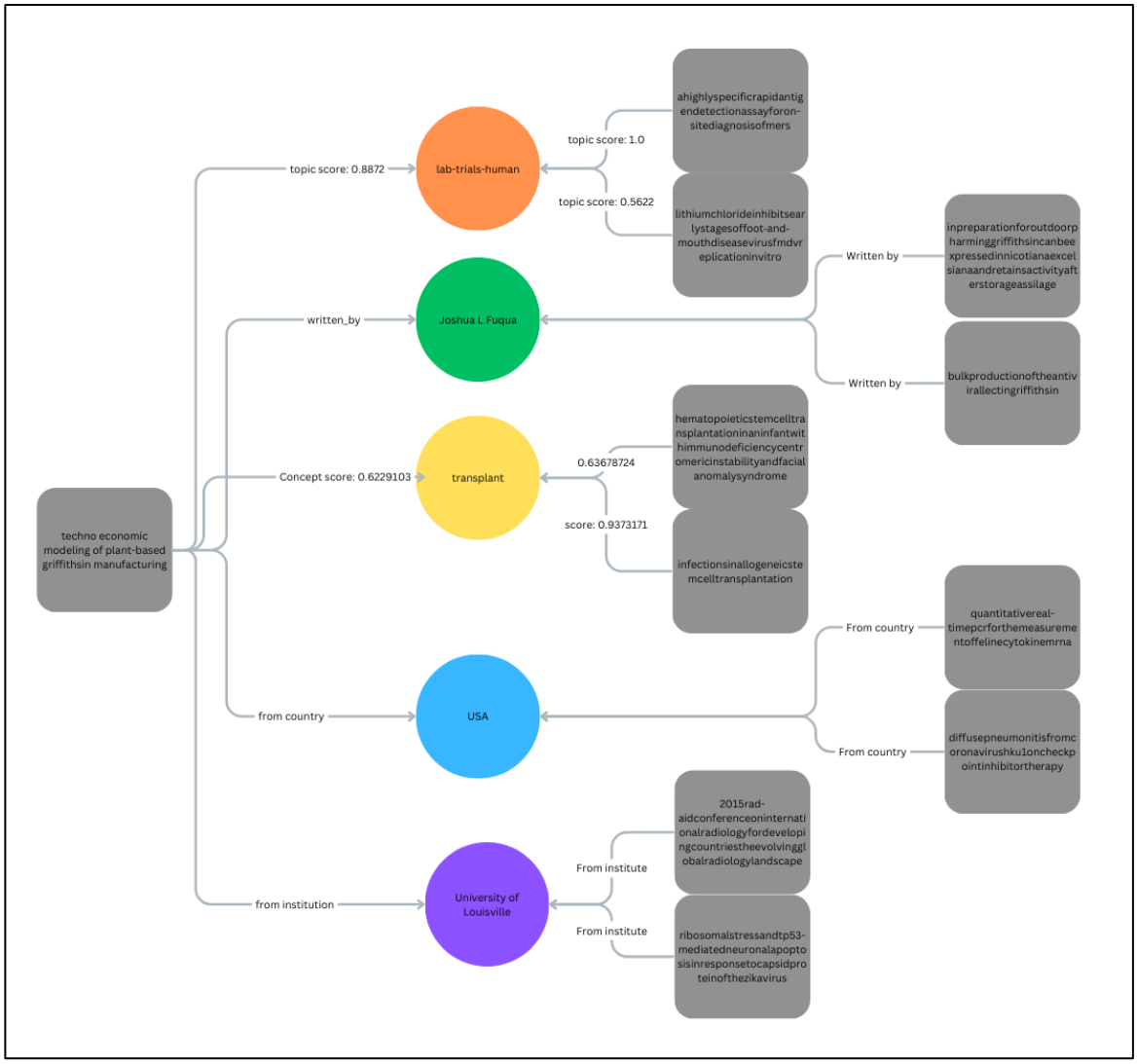


Figure 4-10 Example knowledge graph formation without the sentiment score. We can see that other papers are discovered using the properties of one paper

The visualized knowledge graph centers on the paper *"Technoeconomic Modeling of Plant-Based Griffithsin Manufacturing"* and shows how it connects to key metadata elements such as topic, author, concept, country, and institution. The paper is strongly associated with the topic *lab-trials-human* and the concept *transplant*, supported by relevance scores. It is authored by *Joshua L. Fuqua*, affiliated with the *University of Louisville*, and linked to the *USA* as the publishing country. The graph also shows related papers that share similar topics, concepts, or institutional affiliations, highlighting the interconnected nature of

scientific research and demonstrating how such a structure enables multidimensional exploration and discovery across the COVID-19 literature.

Figure 4-11 Example knowledge graph formation with the sentiment score.

We can clearly see that additional papers are discovered if we include the sentiment score for the example paper.

4.3 Other statistics of the data

Table 4-4 General statistics of the resulting knowledge graphs in the end of the research

Statistic	Value
Country Edge count	453,633
unique_authors	229,911
average_papers_per_author	1.45
unique_topics	10
Topic edge count	129,530
unique_institutions	21,297
concept_count	76,768
concept edges	1,783,589

This research dataset includes many connections between different types of information. The Country Edge Count (453,633) shows how many research papers are linked to different countries, based on where they were published or where the authors are from. Since some papers have authors from multiple countries, they can be connected to more than one place. There are 229,911 unique authors, meaning nearly 230,000 different researchers contributed to the COVID-19 research papers.

On average, each author wrote about 1.45 papers, showing that while some researchers published multiple papers, most only worked on one. The papers are grouped into 10 unique topics, helping to organize them by subject. These topics are connected to papers through 129,530 topic edges, which show how different studies fit into various topics.

In addition to authors and topics, institutions and concepts also play an important role in the dataset. There are 21,297 unique institutions, meaning over 21,000 research organizations contributed to COVID-19 studies. The dataset also includes 76,768 unique concepts, which are keywords or ideas discussed in the papers. Papers and concepts are connected through 1,783,589 concept edges, showing that many papers share common concepts and discuss related topics. These connections help in understanding how different papers relate to each other and make it easier to find important research based on shared ideas or themes.

Table 4-5 Node connectivity

Distinct nodes	Node type	Edges
10	Topic	129,530
21,297	Institution	476,865
229,911	Author	335,138
76,768	Concept	1,783,786
100	Sentiment	32,299
	Total Edges	Average Connectivity Per Node
Before sentiment	2,725,319	8.31
After Sentiment	2,757,618	8.41

The data presented in the table highlights the structure and impact of integrating sentiment analysis into the COVID-19 knowledge graph. Initially, the knowledge graph was constructed using metadata from five key node types: Topic, Institution, Author, Concept, and Sentiment. The Concept node had the highest number of edges (1,783,786), indicating its strong role in linking scientific papers based on shared concepts. Institution and Author nodes also contributed significantly with 476,865 and 335,138 connections, respectively.

After incorporating sentiment scores as a new node type represented by 100 distinct sentiment nodes a total of 32,299 new edges were added. This brought the total number of edges in the graph from 2,725,319 (before sentiment integration) to 2,757,618 (after sentiment integration). Consequently, the average connectivity per node increased from 8.31 to 8.41. This increment demonstrates that including sentiment as an additional dimension enriched the graph's structure by introducing new relational clusters. It enhanced the graph’s ability to uncover hidden connections and improve knowledge retrieval, particularly by linking papers with similar narrative tones or sentiment characteristics.

The programming language used for this research was python. The pandas library is mainly used for the data processing and matplotlib was used for visualization. To create the network-like structure for the visualization of the knowledge graph, networkx package was used.

Chapter 5 Discussion and Conclusions

To create multifaceted knowledge graphs, the study uses information from AWS and a large dataset from the CORD-19 corpus. These graphs allow for the systematic study of COVID-19-related literature by capturing a variety of dimensions, including themes, nations, institutions, authors, concepts, and sentiment scores. It is also noteworthy that sentiment analysis was introduced using BERT-based models since it adds another level of context by classifying papers with comparable sentiment profiles.

The usefulness of knowledge graphs is increased when they are combined into a single structure, even when individual graphs offer insights into attributes like subjects or organizations. Better retrieval capabilities are provided by the combined graph, which makes it easier to find connections that are hidden in separate networks. To navigate the extensive COVID-19 literature, for example, a single query can now return related papers based on a variety of factors, including sentiment scores, similar themes, or even shared authors.

The knowledge graph gains a great deal of value from the addition of sentiment analysis. This study creates novel relational clusters that go beyond conventional metadata by giving papers sentiment scores based on their textual content. The combined graph's analytical depth is enhanced by the new connections created by papers with comparable sentiment scores. This contribution shows how sentiment-based clustering can help researchers generate hypotheses and prioritize their findings by exposing hidden linkages and trends in the literature.

The findings demonstrate how sentiment analysis can be used to improve the integrated knowledge graph. The graph only uses predetermined attributes, like subjects or institutions, in the absence of sentiment analysis. On the other hand, adding sentiment scores adds more levels of linkage. It makes it possible, for example, to find publications with similar narrative tones or broad conclusions, which could point to agreement or new areas of interest within the academic community.

These findings align directly with the research objective of building a multifaceted knowledge graph capable of semantically organizing and navigating COVID-19 literature. By capturing various dimensions—including countries, topics, institutions, authors, and

concepts the graph serves as a robust foundation for exploring relationships and uncovering latent patterns across the dataset.

The integration of sentiment analysis using a BERT-based classifier significantly enhances the utility of the knowledge graph. This represents a novel extension of traditional metadata-based graphs by introducing relational clustering based on textual sentiment. The sentiment-based graph added 32,299 new edges, demonstrating the added analytical depth and indicating how papers with similar emotional or rhetorical tone may be connected even if they differ by topic or institution. This directly supports the secondary research aim of exploring how sentiment signals can augment traditional knowledge discovery.

From a practical standpoint, the enhanced knowledge graph facilitates multi-dimensional information retrieval. Researchers can now query not only by keywords, authors, or institutions but also by sentiment orientation a valuable feature for identifying consensus in literature, emerging concerns, or conflicting views. For instance, public health experts might retrieve clusters of papers with negative sentiment to better understand community-level challenges, while policymakers might focus on positively scored clusters to extract actionable strategies.

Moreover, the unified graph supports hypothesis generation by allowing researchers to trace semantic, institutional, and emotional linkages across disparate documents. This is especially useful in fast-moving domains like pandemic response, where understanding shifts in sentiment across time and topics can inform adaptive policy design and future preparedness strategies.

While the sentiment component adds interpretive power, it also introduces potential sources of error. The sentiment scores were generated using a pre-trained BERT model, fine-tuned for document-level classification. However, several limitations apply:

- Domain mismatch: BERT models trained on general corpora may not fully capture the nuanced and technical language found in scientific literature, particularly in biomedical texts.

- Contextual ambiguity: Sentiment in academic writing tends to be subtle and context-specific. For example, a study describing negative outcomes may not

reflect pessimism but scientific objectivity.

- Imbalanced data: If the sentiment classifier is not trained on a balanced set of positive, negative, and neutral scientific texts, it may skew toward the dominant class, reducing the reliability of minority sentiment detection.

To mitigate these issues, future work should include fine-tuning the model on a labelled COVID-19 scientific corpus, employing cross-validation to measure accuracy, and conducting human-in-the-loop evaluations for a subset of predictions to assess alignment with expert interpretation. A confidence threshold could also be introduced to flag uncertain classifications.

Comprehensive statistics that highlight the scope of the created graphs were obtained from the analysis. For instance, the sentiment knowledge graph by itself added 32,299 connections, while the composite graph showed complex relationships between 476,865 institutional edges, 178,3589 idea edges, and other edges. These indicators demonstrate the methodology's stability and scalability when working with big datasets.

The possibilities of knowledge graphs as a tool for organizing and navigating enormous collections of scientific literature are highlighted by this study. When sentiment analysis is added, the combined knowledge graph shows exceptional ability in revealing hidden relationships, seeing patterns, and promoting effective information retrieval. The methodology establishes a standard for augmenting the analytical capabilities of knowledge graphs in tackling extensive issues such as the COVID-19 pandemic by connecting sentiment-driven insights with conventional information. To further increase its usefulness, future research might apply this strategy to other fields and use other machine learning methods and data modalities.

The pandemic of COVID-19 has been very challenging year for humanity. Searching through relevant scientific papers and literature to find the correct one is of utmost importance. This research focused on leveraging knowledge graphs to structure and analyse the vast corpus of COVID-19-related literature. A knowledge graph is a way to describe semantical relationships between entities. We can create knowledge that connects multiple papers by acquiring metadata about the CORD-19 papers from the AWS website.

By constructing multiple knowledge graphs based on distinct properties such as country, authors, topics, institutions, concepts, and sentiment scores, the approach enabled a multi-dimensional exploration of the dataset. Using a BERT model for Sentiment classification, we were able to add additional layer of insight. This sentiment score allowed us to cluster similar papers together based on score similarity. The integration of these graphs supports efficient information retrieval and uncovers hidden patterns across the CORD-19 dataset. While this is an initial step, the methodology provides a strong foundation for future enhancements in knowledge graph applications, particularly in handling large and diverse datasets.

Overall, the construction of this multi-layered knowledge graph augmented with sentiment-based edges demonstrates a scalable, data-rich approach to navigating complex scientific literature. It not only meets the initial research objectives but also lays the groundwork for real-world applications in literature review, research prioritization, and policymaking. However, to ensure reliable insights, especially from sentiment analysis, further validation and refinement of the sentiment model are essential. This work sets a precedent for how semantic enrichment and emotional context can be meaningfully combined in knowledge graphs to address pressing global challenges like COVID-19.

Chapter 6 References

[1] J. Bosman, "www.nytimes.com," 07 2020. [Online]. Available: https://www.nytimes.com/2020/07/27/us/coronavirus-data.html.

[2] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh and D. Batra, "Embodied question answering," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[3] Q. a. L. M. a. W. X. a. P. N. a. H. G. a. M. J. a. T. J. a. L. Y. a. Z. H. a. L. W. a. o. Wang, "COVID-19 literature knowledge graph construction and drug repurposing report generation," *arXiv preprint arXiv:2007.00576,* 2020.

[4] Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine*.

[5] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, *395*(10223), 497-506.

[6] W. a. F. R. a. o. McKibbin, "The economic impact of COVID-19," *Economics in the Time of COVID-19,* vol. 45, 2020.

[7] Morawska, L., & Milton, D. K. (2020). It is time to address airborne transmission of COVID-19. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, ciaa939.

[8] Prather, K. A., Wang, C. C., & Schooley, R. T. (2020). Reducing transmission of SARS-CoV-2. Science, 368(6498), 1422-1424.

[9] Else, H. (2020). Covid in papers. Nature, 588(24/31), 553.

[10] Patel, J. A., Nielsen, F. B. H., Badiani, A. A., Assi, S., Unadkat, V. A., Patel, B., ... & Wardle, H. (2020). Poverty, inequality and COVID-19: the forgotten vulnerable. Public health, 183, 110.

[11] Pfefferbaum, B., & North, C. S. (2020). Mental health and the Covid-19 pandemic. New England journal of medicine, 383(6), 510-512.

[12] Singhal, A. (2012). Introducing the knowledge graph: things, not strings. Official google blog, 5(16), 3.

[13] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (Csur), 54(4), 1-37.

[14] Li, P., Zhao, Q., Liu, Y., Zhong, C., Wang, J., & Lyu, Z. (2024). Survey and Prospect for Applying Knowledge Graph in Enterprise Risk Management. Computers, Materials & Continua, 78(3).

[15] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... & Welty, C. (2010). Building Watson: An overview of the DeepQA project. AI magazine, 31(3), 59-79.

[16] Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. Complex & Intelligent Systems, 7(1), 439-457.

[17] COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation

[18] Piplai, A., Mittal, S., Joshi, A., Finin, T., Holt, J., & Zak, R. (2020). Creating cybersecurity knowledge graphs from malware after action reports. IEEE Access, 8, 211691-211703.

[19] Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2006). Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer Science & Business Media.

[20] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.

[21] Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., & Kodamullil, A. T. (2021). COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics, 37(9), 1332-1334.

[22] Wang, X., Song, X., Li, B., Guan, Y., & Han, J. (2020). Comprehensive named entity recognition on cord-19 with distant or weak supervision. arXiv preprint arXiv:2003.12218.

[23] Kejriwal, M. (2020). Knowledge graphs and COVID-19: opportunities, challenges, and implementation. Harv. Data Sci. Rev, 11, 300.

[24] Hope, T., Amini, A., Wadden, D., van Zuylen, M., Parasa, S., Horvitz, E., ... & Hajishirzi, H. (2020). Extracting a knowledge base of mechanisms from COVID-19 papers. arXiv preprint arXiv:2010.03824.

[25] Reese J, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Fontana T, Blau H, Matentzoglu N, Harris NL, Munoz-Torres MC, Robinson PN, Joachimiak MP, Mungall CJ. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. bioRxiv [Preprint]. 2020 Aug 18:2020.08.17.254839. doi: 10.1101/2020.08.17.254839. Update in: Patterns (N Y). 2021 Jan 8;2(1):100155. doi: 10.1016/j.patter.2020.100155. PMID: 32839776; PMCID: PMC7444288.

[26] 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:2007.00576.

[27] 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:2007.12731.

[28] Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., ... & Kodamullil, A. T. (2021). COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics, 37(9), 1332-1334.

[29] Lei, Z., Sun, Y., Nanehkaran, Y. A., Yang, S., Islam, M. S., Lei, H., & Zhang, D. (2020). A novel data-driven robust framework based on machine learning and knowledge graph for disease classification. Future Generation Computer Systems, 102, 534-548.

[30] Sharma, K., Zhang, Y., Ferrara, E., & Liu, Y. (2021, August). Identifying coordinated accounts on social media through hidden influence and group behaviours. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 1441-1451).

[31] Huang, W., Liu, J., Li, T., Ji, S., Wang, D., & Huang, T. (2022). Fedcke: Cross-domain knowledge graph embedding in federated learning. IEEE Transactions on Big Data, 9(3), 792-804.

[32] Liu, B., Fang, Y., Wang, X., & Li, X. (2024, August). Federated Knowledge Graph Embedding Unlearning via Diffusion Model. In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data (pp. 272-286). Singapore: Springer Nature Singapore.

[33] Allen Institute for AI. (n.d.). COVID-19 Open Research Dataset (CORD-19). AWS Marketplace. Retrieved December 22, 2024, from https://aws.amazon.com/marketplace/pp/prodview-ybwpxcqlznbas#offers

[34] Reese J, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Fontana T, Blau H, Matentzoglu N, Harris NL, Munoz-Torres MC, Robinson PN, Joachimiak MP, Mungall CJ. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. bioRxiv [Preprint]. 2020 Aug 18:2020.08.17.254839. doi: 10.1101/2020.08.17.254839. Update in: Patterns (N Y). 2021 Jan 8;2(1):100155. doi: 10.1016/j.patter.2020.100155. PMID: 32839776; PMCID: PMC7444288.

[35] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2).

[36] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[37] Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588.

[38] Wang, H., Du, H., Qi, G., Chen, H., Hu, W., & Chen, Z. (2022). Construction of a Linked Data Set of COVID-19 Knowledge Graphs: Development and Applications. *JMIR medical informatics*, *10*(5), e37215. https://doi.org/10.2196/37215

[39] Wang et al., NLP-COVID19 2020).CORD-19: The COVID-19 Open Research Dataset (https://aclanthology.org/2020.nlpcovid19-acl.1/)

[40] Rahdari, Behnam, et al. "CovEx: An exploratory search system for COVID-19 scientific literature." *University of Pittsburgh* (2020).

[41] Wang, Haofen et al. “Construction of a Linked Data Set of COVID-19 Knowledge Graphs: Development and Applications.” *JMIR medical informatics* vol. 10,5 e37215. 13 May. 2022, doi:10.2196/37215