

LB/TH/41/2025

TH5996

**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 COVID-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 COVID-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

Abstract	ii
Declaration	iii
Acknowledgment	iv
Table of Contents.....	v
List of Figures.....	vii
List of Tables.....	viii
Chapter 1 Introduction	1
1.1 Research Problem.....	2
1.2 Research Objective.....	3
1.3 Organization of the report.....	3
Chapter 2 Literature Review	4
2.1 Covid-19 A global pandemic.....	4
2.1.1 Introduction and Background.....	4
2.1.2 Clinical Presentation and Symptoms.....	4
2.1.3 Interdisciplinary Research Efforts.....	5
2.1.4 Challenges in Research and Knowledge Synthesis.....	6
2.1.5 Impacts Beyond Health.....	6
2.2 Application of Knowledge graph in general.....	7
2.2.1 Introduction to Knowledge Graphs.....	7
2.2.2 Challenges in Knowledge Graph Construction.....	7
2.2.3 Knowledge Graphs in Structured Knowledge Representation.....	8
2.2.4 Applications in Key Domains.....	8
2.2.4.1. Question Answering Systems.....	8
2.2.4.2. Recommender Systems.....	9
2.2.4.3. Information Retrieval Systems.....	9
2.2.4.4. Domain-Specific Applications.....	9
2.2.5 Knowledge Graph Embeddings.....	10
2.2.6 Knowledge Graphs and Machine Learning.....	10
2.2.7 Gaps in Existing COVID-19 Knowledge Graphs.....	10
2.2.8 Future Directions.....	12
2.3 Application of Knowledge graph on COVID-19 literature.....	12
2.3.1 Knowledge Graph Construction for COVID-19 Research.....	13
2.3.3 Comparative Overview of Prominent COVID-19 Knowledge Graphs.....	14

2.3.3 Applications Beyond Literature Search	15
2.3.4 Recent Advances in Knowledge Graph Techniques	16
Chapter 3 Methodology	18
3.1 Knowledge graph Summary.....	18
3.2 Data Collection and Dataset Preparation	19
3.2.1 The CORON-19 dataset.....	19
3.2.2 Metadata Extraction from AWS.....	20
3.2.3 Full Dataset from Kaggle.....	22
3.3 Knowledge Graph Construction.....	23
3.3.1 Properties of the Knowledge Graph.....	23
3.3.2 Knowledge graph creation	24
3.3.3 Understanding BERT and Its Role in Text Classification	28
3.3.3.1 Key Features of BERT	28
3.3.3.2 BERT in Text Classification	29
3.3.3.3 BERT's Strengths in Sentiment Classification	30
3.3.4 Sentiment Score	31
3.3.4.1 Sentiment Model Selection	31
3.3.4.2 Sentiment Score Calculation	31
3.3.4.3 Example Calculation	32
3.3.5 Evaluation of Knowledge Graph Effectiveness	33
Chapter 4 Analysis and Results.....	34
4.1 Visualization of Knowledge graphs.....	35
4.1.1 Topic Knowledge Graph.....	35
4.1.2 Country Knowledge Graph	37
4.1.3 Author Knowledge Graph.....	38
4.1.4 Concept Knowledge Graph.....	39
4.1.5 Institution Knowledge Graph.....	40
4.1.6 Sentiment Knowledge Graph.....	40
4.3 Other statistics of the data.....	45
Chapter 5 Discussion and Conclusions.....	47
Chapter 6 References	51

List of Figures

Figure 1-1 Sample Knowledge graph created using COVID -19 data (Molecular relationship) https://covid19.tubitak.gov.tr/en/covid-19-bilgi-grafikleri/ ,.....	2
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/)	13
Figure 3-1 Example of a knowledge Graph equivalency.....	18
<i>Figure 3-2 Example of Paper node</i>	21
Figure 3-3 Sample of Concept Node.....	21
Figure 3-4 Sample relationship of paper to concept node.	22
Figure 3-5 Sample of a single knowledge graph	26
Figure 3-6 Random 10 papers with the concept it is associated with.....	27
Figure 3-7 Combining all the knowledge graph into one. [25].....	27
Figure 3-8 Transformer Architecture [36]	28
Figure 4-1 A paper connect with different topics	35
Figure 4-2 Topics which connect multiple papers together	36
Figure 4-3 Multiple papers connection shown which seemed to originate from the same country.	37
Figure 4-4 Paper shown with all its authors.....	38
Figure 4-5 multiple papers with connected authors	38
Figure 4-6 Papers connected to concepts with their associate confidence scores.	39
Figure 4-7 Papers connected with many Institutions.	40
Figure 4-8 Other papers which can be discovered with the addition of the sentiment scores for all the papers.....	41
Figure 4-9 A combined knowledge graph with all the properties mentioned in this research.	42
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	43
Figure 4-11 Example knowledge graph formation with the sentiment score.	44

List of Tables

Table 2-1 Comparative Overview of Prominent COVID-19 Knowledge Graphs.....	14
Table 3-1 First 10 fields of the Topics Knowledge graph file	25
Table 4-1 Number of lines in the available files from the AWS dataset	34
Table 4-2 Number of connections made after the formation of the knowledge graphs.....	35
Table 4-3 Color for each of the property in the combined knowledge graph.....	42
Table 4-4 General statistics of the resulting knowledge graphs in the end of the research	45
Table 4-5 Node connectivity.....	46