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