Ryerson University
Data Science & Analytics Graduate Studies

2021-2022 M.Sc. in Data Science and Analytics

Cyber Resilience Topic Modeling

Bhakti Bhatt **Supervisor: Dr. Atefeh Mashatan**

Objective

Topic modelling is a key natural language processing method used to uncover hidden insights from large collections of documents. It is an unsupervised machine learning technique that scans a set of documents, detects word patterns, and clusters similar word groups to characterize a set of documents into “topics”. This research uses a probabilistic topic modelling algorithm known as Latent Dirichlet Allocation (LDA) which is widely used to identify hidden semantic structure in a large text corpus. In this study, experiments were conducted to realize LDA implementation to perform topic modelling on academic literature related to cyber resilience. Cyber resilience has evolved into an essential concept for research and discussion in academic literature. This inspires the use of LDA to assist in analyzing and extracting the semantic structure of the corpus. This study identifies the key topics in cyber resilience, exhibits topic distributions over time, and evaluates the impact of COVID-19 on research areas in the cyber resilience landscape.

Background

It was critical to explore and draw value from the existing literature on topic modelling, cyber resilience, and the implications of COVID-19 on cyber resilience.

Topic Modelling

The current professional literature provides insights on a variety of areas where topic modelling is valuable and validates its significance in research today. Its unique ability to take unstructured text data, transform, and cluster it into meaningful topics helps researchers gain insights into historical trends and emerging themes within any field of study. As such, it is an advantageous tool to use for the purpose of this research. The papers with the highest degree of similarity to our study employ topic modelling with LDA techniques to gain insights on cyber security. Kolini et al., Alagheband et al., Kumar et al. all employ topic modelling to conduct different analysis on the broader scope of cyber security. This study narrows the focus from cyber security to cyber resilience. As well, articles in this study are limited to academic literature.

Cyber Resilience

There are a noticeable number of research papers on cyber resilience that describe the importance, methodology, improvements, and metrics, which fuels our motivation to further dissect the multidimensional framework and conduct a topic analysis. This research describes the main topics related to cyber resilience that are currently discussed in academic literature.

Implications of COVID-19 on Cyber Resilience

The unprecedented worldwide effect we faced as a society as a result of the COVID-19 pandemic, has trigged novel research across many industries and sectors. Various articles discuss the need for enhancing and developing cyber security and resiliency post-pandemic in IT systems, supply chains, healthcare, businesses, and financial institutions. Given the noticeable amount of academic literature triggered by circumstances originating from COVID-19, our study also explores the resulting impact on research surrounding cyber resilience.

To our best knowledge, existing research has not topically explored cyber resilience literature through an LDA model, or the change in landscape in cyber resilience as a result of the pandemic, making our study the very first to add a new dimension.

Methodology

The figure below depicts the overall methodology employed in this study. There are four primary phases undertaken to achieve the results in this study. The first phase covers an initial analysis and retrieval of documents related to cyber resilience. The second phase includes preprocessing the text, using techniques such as stopword removal, special characters removal, tokenization and lemmatization to finalize the corpus. Subsequently, we analyze the corpus using two complementary algorithms, topic modelling and bibliometric analysis. For topic modelling, we use the LDA model which reveals the themes present in research and the underlying trends. The bibliometric analysis helps understand the corpus from an alternative perspective. As part of the bibliometric analysis, we explore co-authorship and co-occurrences of key words to holistically depict the state of research. Finally, we conclude by processing and analyzing the results from separate LDA models for pre-pandemic and post-pandemic datasets.

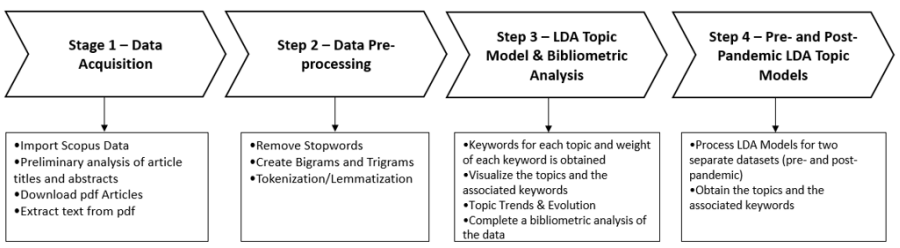


Figure 1. Methodology Employed for this study.

- LDA is a type of generative probabilistic algorithm that treats each document as a bag of words and aims to identify the latent themes in the papers. It requires three primary inputs: a corpus (term-document frequency matrix), a dictionary (id2term matrix), and number of topics. We used the coherence score metric to determine the optimal number of topics. The coherence score of a model is the degree of similarity between words in a topic; providing how well the words support each other. The model with the highest coherence score generally best represents the full list of documents.
- Iterative experiments were conducted to find the optimal number of topics. The LDA model was processed against different numbers of topics (e.g., 1 to 50) for several iterations and the coherence scores were recorded for each model. Then, the average coherence score was calculated for each. The optimal number of topics was selected based on a combination of the maximum coherence score and qualitative assessment of the output. The output of the LDA model was the topic distribution among the documents and the keyword composition.
- In this study, the topic labels were defined based on manual examinations of the keywords and the most representative documents in topic, as done in other similar studies.
- Bibliometric analysis was performed using the VOSViewer platform. We analyzed keyword networks and themes for abstracts, titles, author keywords, and co-authorship amongst countries. Extracting insights from the bibliometric analysis ensured our preparedness for the manual task of topic labelling.

Results

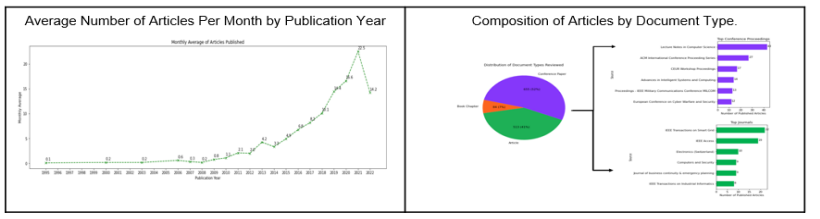


Figure 2. Results obtained from the exploratory data analysis.

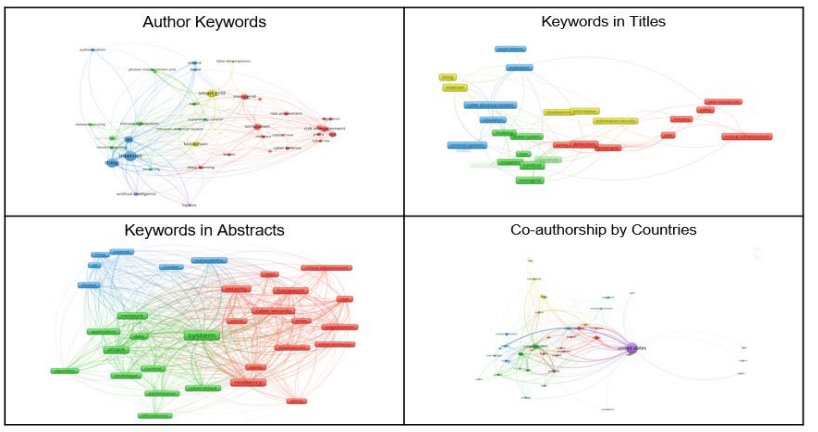


Figure 3. Results obtained from the bibliometric analysis using VOSViewer.

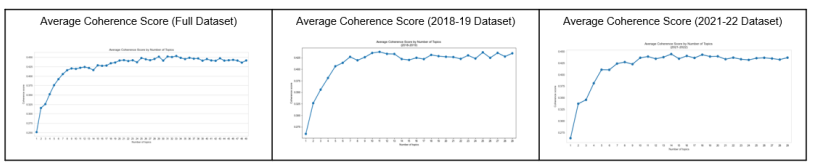


Figure 4. Average coherence scores obtained in different datasets for different number of topics.

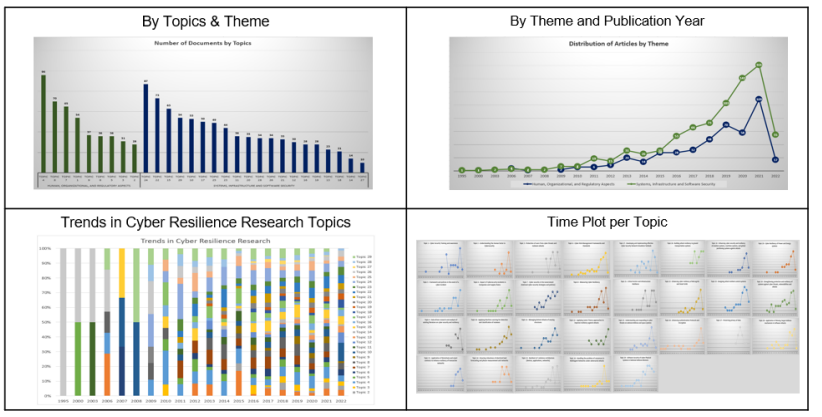
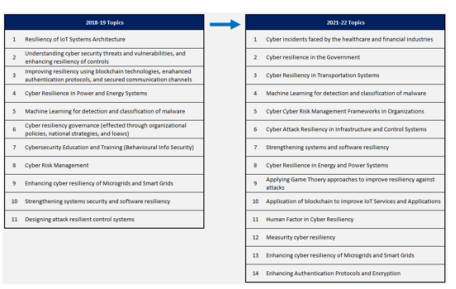


Figure 5. Results obtained from the topic modelling on the full corpus.



2018-19 Topics	2021-22 Topics
1 Resiliency of IoT Systems Architecture	1 Cyber incidents faced by the healthcare and financial industries
2 Understanding cyber security threats and vulnerabilities, and enhancing resiliency of controls	2 Cyber resilience in the Government
3 Improving resiliency using blockchain technologies, enhanced authentication protocols, and secured communication channels	3 Cyber Resiliency in Transportation Systems
4 Cyber Resilience in Power and Energy Systems	4 Machine Learning for detection and classification of malware
5 Machine Learning for detection and classification of malware	5 Cyber Cyber Risk Management Frameworks in Organizations
6 Cyber resiliency governance (effected through organizational policies, national strategies, and loans)	6 Cyber Attack Resiliency in Infrastructure and Control Systems
7 Cybersecurity Education and Training (Behavioural Info Security)	7 Strengthening systems and software resiliency
8 Cyber Risk Management	8 Cyber Resilience in Energy and Power Systems
9 Enhancing cyber resiliency of Microgrids and Smart Grids	9 Applying Game Theory approaches to improve resiliency against attacks
10 Strengthening systems security and software resiliency	10 Application of blockchain to improve IoT Services and Applications
11 Designing attack resilient control systems	11 Human Factor in Cyber Resiliency
	12 Measurity cyber resiliency
	13 Enhancing cyber resiliency of Microgrids and Smart Grids
	14 Enhancing Authentication Protocols and Encryption

Figure 6. Results obtained from pre-pandemic and post-pandemic topic models.

Conclusions

- Topic Modelling using the LDA Algorithm can be used to identify hidden semantic structure through unsupervised learning in a large text corpus. In this study, experiments were conducted to realize an LDA implementation to perform topic modelling on textual data.
- A dataset containing full text extracted from 1252 articles on cyber resilience was used and a LDA model was trained to identify the distribution of topics throughout the text corpus. We evaluate the results from the model both qualitatively and quantitatively.
- The bibliometric analysis highlighted areas such as artificial intelligence, smart city, risk assessment, simulation, and botnet, which can be considered novel research opportunities that are arising.
- Based on the statistical distribution of documents in each of the topics obtained through our pre-pandemic and post-pandemic LDA models, we report that post-pandemic, an increase was observed in research on cyber incidents in healthcare and financial sectors, and cyber security education and training. Meanwhile, several topics remained important in perpetuity, such as the use of machine learning for the detection and classification of malware, and cyber risk management.
- The discovered topics and themes with their respective trends highlight future research opportunities in the domain, which can be used by academicians and industry professionals to make informed decisions.