A Global Supply Chain Risk Management Framework: An Application of Text-mining to Identify Region-specific Supply Chain Risks

Chih-Yuan Chu[a], Kijung Park[b] and Gül E. Kremer[a]*

[a] *Department of Industrial and Manufacturing Systems Engineering, Iowa State University, 2529 Union Drive, Ames, IA 50011, USA*
[b] *Department of Industrial and Management Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Korea*

Abstract

Nowadays global supply chains enable companies to enhance competitive advantages, increase manufacturing flexibility and reduce costs through a broader selection of suppliers. Despite these benefits, however, insufficient understanding of uncertain regional differences and changes often increases risks in supply chain operations and even leads to a complete disruption of a supply chain. This paper addresses this issue by proposing a text-mining based global supply chain risk management framework involving two phases. First, the extant literature about global supply chain risks was collected and analyzed using a text-based approaches, including term frequency, correlation, and bi-gram analysis. The results of these analyses revealed whether the term-related content is important in the studied literature, and correlated topic model clustering further assisted in defining potential supply chain risk factors. A risk categorization (hierarchy) containing a total of seven global supply chain risk types and underlying risk factors was developed based on the results. In the second phase, utilizing these risk factors, sentiment analysis was conducted on online news articles, selected according to the specific type of risk, to recognize the pattern of risk variation. The risk hierarchy and sentiment analysis results can improve the understanding of regional global supply chain risks and provide guidance in supplier selection.

Keywords: Global supply chain; Risk management; Data analytics; Text mining; Sentiment analysis; Google News

1. Introduction

Global supply chains have been widely discussed to boost competitive advantages [1]. From an economic perspective, global supply chain management can create a win-win situation for associated stakeholders in that every role (or node) in the global supply chain network can obtain operational benefits when all stages related to the product/component production (e.g., design, manufacturing, assembly, testing, and marketing) is coordinated worldwide [2]. Although operating global supply chains has

been considered as a common profit advantage in most industries, uncertain regional factors regarding global business and logistics such as government stability, insufficient infrastructure in a country, and trade barriers may make it difficult to have successful supply chain management [3, 4].

A global supply chain network can be interpreted as an integration of both internal and external stakeholders. This integration has positive impacts on corporate operations such as innovative strategy, product quality and profitability [5]. However, critical risks and challenges in global business operations may lead to the disruption of the global supply chain although suppliers in multiple tiers and various nations are fully coordinated to decrease the total supply chain cost. Christopher et al. [6] proposed a classification for global sourcing risks that include supply risk, environmental and sustainability risk, process and control risk, and demand risk. Among these risks, supply risk, process and control risk, and demand risk have been widely discussed in practice and academia because they are relatively easier to evaluate through quantitative methods common to domestic supply chain cases. From a global business dynamics vantage point, on the other hand, environmental and sustainability risks such as fluctuations in environmental protection legislations, taxation differences, and political stability are difficult to analyze and evaluate due to related uncertainties affected by local political, social and economic realities. This insufficient understanding of regional – often across national boundaries – differences and thus changes may significantly influence and disrupt the entire supply chain performance [6]. One significant current challenge in supply chain risk management is still about modeling and quantifying the risks [7]. A comprehensive risk categorization is needed to better understand the scope of global supply chain risk factors due to regional differences.

The above mentioned issues lead to the research questions: 1) What are the significant global supply chain risks due to regional differences? 2) How can such risks be identified and organized objectively? 3) How can the risk variation (factors and levels) patterns be recognized? Although there are previous studies on global supply chain risk management, most research applied case studies, operational-modelling or probability-related methodologies to categorize risks and address supply chain design problems [8, 9, 10]. Moreover, studies using qualitative literature review methods to classify global supply chain risks have challenges in validating due to their limited sample [11]. Therefore, a data-driven method is needed for researchers and practitioners to identify global supply chain risks and understand the significant risk patterns [12].

Artificial intelligence has been increasingly applied in supply chain risk management area; nevertheless, very few of these studies incorporated text analytics to

extract useful information on supply chain risks [13]. To identify regional risks of a global supply chain by a data-driven approach, input data collection is the first and most important step to undertake. However, there is a great deal of text information on geographical differences and supply chain risks in published articles, business magazines, and news articles. General public is also able to receive information about natural disasters, government policies, and company operations from news media. Therefore, this study attempts to discover how this information can be exploited to reveal signals on global supply chain risks. Nowadays, such text-based information could be easily accessed through open portals or databases for academic purposes. Research databases such as Scopus, Engineering Village, and Google Scholar have a wide variety of peer-reviewed research articles, which are organized for retrieval by search algorithms. Published news articles are also stored on publisher's website or database. Platforms such as Google News or Yahoo News collect the latest news articles from different mass communication media. These exemplify the abundance of information and sharing mechanisms, and demonstrates that online news media have already become a key element of social, economic, and cultural life worldwide [14]. This trend presents the possibility that such information can be sources of input to this research; peer-reviewed journal articles can help define important risk factors with their careful review processes, and online news articles aid in risk variation pattern recognition because of their content and time-specific relevance.

To analyze these articles directly for discovering valuable insights about regional supply chain risks, a text-oriented, unstructured data analysis method is required. Therefore, text mining is adopted to extract information from the input documents. Text mining is an approach that can analyze a large amount of documents (in a form of pre-processed text format) and extract valuable information. Text mining techniques have been applied to numerous areas such as political document analysis, energy and service industries, information technology, and healthcare service sectors due to their outstanding capability in analyzing unstructured data [15]. It is extensively applied in knowledge discovery, retrieval, and management. Artificial intelligence is also implemented in more advanced text-mining tools to analyze textual data to discover patterns and insights in unstructured texts [10].

This study addresses the above stated research questions by developing a text-mining-based risk management framework for global supply chains that consider up-to-date text information. This framework could provide guidance for companies to develop a resilient global supply chain considering the influences of regional differences and changes.

The remainder of this paper is organized as follows. Section 2 discusses previous studies of global supply chain risk management and data analytics (including text mining) applications. Section 3 demonstrates the proposed research framework. Section 4 presents the implementation and results of defining potential risk factors and risk variation pattern recognition incorporating sentiment analysis. A discussion on relevance to previous research and limitations of this study is also presented in this section. Finally, the summary of the proposed framework, highlights of contributions, and future research directions are included in Section 6.

2. Literature review

In this section, previous studies were investigated in two categories. First category reviews the studies that discussed risk management in the context of global supply chain. The second one illustrates the development of data analytics and its applications in supply chain management. At the end of this section, a brief summary pointing to the research gaps is presented to highlight the importance of the research question tackled herein.

2.1. Global supply chain risk management

Due to the improvement of information sharing technologies and the trend of globalization, domestic supply chains of most enterprises have evolved into global supply chains by integrating multi-nation-resources to their current operations [16]. In spite of the advantages of global supply chains, such as access to new markets, capability to reduce cost, etc., there are numerous potential issues that remain to be overcome by practitioners and scholars. In comparison to a relatively simple domestic supply chain, the complexity of global sourcing frequently results from additional regional risks such as taxes, export regulations, exchange rates, etc. [17]. Moreover, diversity in culture, languages, and political environment are among the reasons why a global supply chain is incredibly difficult to operate at a high performance level [16]. Therefore, understanding regional risks is crucial for region selection process in global supply chain design.

Christopher et al. [6] classified global supply chain risks as supply risk, environmental and sustainability risk, process and control risk, and demand risk. The environmental and sustainability risks contain political instability, export regulations, tariffs, etc., which would have negative impacts on economic, social, and sustainability aspects of an enterprise. Ho et al. [18] suggested that global supply chain risks can be categorized into macro and micro risks. Macro risks comprise factors such as natural disasters, political environment, and regional regulation; while micro risks refer to

demand risk, supply risk, manufacturing risk, and infrastructure risk [19, 20, 21].

Risk assessment methods have also been discussed. Aloini et al. [9] proposed risk identification and risk quantification. The risk identification involves factors that would lead to potential disruption of the network (i.e., risk factors) and their possible effects on performance should be determined. For their assessment, risk factors should also be prioritized quantitatively according to their occurrence probabilities, severity, and detection difficulties. For example, Tsai et al. [22] adopted the analytic hierarchy process (AHP) to implement risk calibration on outsourcing retail chains in Taiwan. Trkman and McCormack [23] built a risk identification conceptual model based on supplier attributes, performance, and the specific environment of operation. Kayis and Karningsih [24] proposed a supply chain risk identification system constructed under a knowledge-based system tool technology.

Although most of the above risk identification models demonstrate strengths on specific cases or types of risk, they were mostly qualitative, with complicated processes and not able to associate risk factors to cost implications. Therefore, a data-driven approach that can systematically deal with unstructured data (e.g., text data) is required to improve global supply chain risk management issues.

2.2. Data analytics applications in supply chain management

Data analytics has become increasingly crucial in numerous fields during the last decade. With the improvement of information communication technologies, data can be accessed and analyzed in a more transparent fashion. Data-driven decisions enable enterprises to manage operations based on evidence instead of speculation or intuition [25]. There are companies that started manipulating and extracting information from data such as Google and Amazon, but the potential benefits of applying data analytics in other industries might be even greater. One of the proofs is the phenomenon that more and more organizations are training and recruiting data scientists to deal with data related to transactions, operations, human resources, etc. [26, 27]. When decision making is no longer about intuition but evidence-based solution, and with this concept flourishing in most of the companies, the positive relation between data science and supply chain management could easily be anticipated.

In data science for supply chain management, the key is to understand the data you are collecting. The categories of data can be listed as sales, consumer, inventory, and location and time [28]. This categorization considers not only the type of data, but also the volume, update frequency, and source variety of data. Although there are different characteristics of data in supply chain management, analytics applications can extract

useful information and create competitive advantages throughout supply chain: from location-based marketing to distribution and logistics to supplier risk management [29].

Data analytics has been applied to supply chain management field considering various operations, such as procurement, manufacturing, warehousing, logistics, and demand management. Nguyen et al. [30] published a review article discussing those applications and the adopted analytics methods to analyze the trends of big data analysis in supply chain management. The article demonstrates how methods such as optimization, classification, simulation, and clustering were implemented to help improve supply chain management. For instances, Choi et al. [31] developed a novel approach by using fuzzy cognitive map to solve supplier selection problems in IT service procurement; Huang and Handfield [32] adopted supply maturity model and information from ERP vendors' public information (e.g., news articles, company announcement, and white papers) to deal with sourcing risk problems; Wu et al. [33] aggregated various data from Taiwanese light-emitting diode industry and used fuzzy and grey Delphi methods to investigate supply chain risks and uncertainties. The above summarized studies demonstrate the potential of data analytics in supply chain issues.

2.3. Text mining and its applications in supply chain management

From the review of the above mentioned previous studies, it is deduced that global supply chain risk management has been discussed as an important issue in academia and industry. The need for data-driven methods has also been emphasized by numerous applications of data analytics in the supply chain field. However, while dealing with risk management problems, a variety of sources related to specific risks involve mostly text data from social media and online news, for example, on natural disasters, the political environment, or tax regulations. This subcategory of data analytics is referred to as text mining.

Text mining is a data analysis method to analyze unstructured text data and extract useful information. Both descriptive analysis and predictive analysis are applicable in text mining. The subtasks of text mining consist of information extraction, text categorization and clustering, visualization, and association rule mining [15]. For example, Chiu and Lin [34] used an information extraction technique to elicit key concepts from product reviews, and integrated them with Kansei Engineering to build a data-driven design application interface. The domain ontology is also a relatively new text mining technique. The purpose of domain ontology is to build a semantic model that captures the common knowledge in the field of interest [35]. One of the recent applications is by Chang and Chen [36], who integrated text analytics and domain ontology to assist in product concept evaluation and selection in crowdsourcing. Text

mining has also been applied to patent knowledge retrieval and extraction, known as patent mining. Trappey et al. [37] incorporated ontology, patent analysis, and metadata analysis in dental implant technology evaluation. As well for the patent mining focus, Govindarajan et al. [38] developed the Excessive Topic Generation, a novel topic modeling approach, which added the word distance relationship feature to the traditional topic generation technique. These previous studies attest to the growing importance and applicability of text mining over the last decade, due to its capability of analyzing large datasets and thereby enhancing artificial intelligence in decision making.

Supply chain management is another research area that text mining is utilized more and more frequently. For example, Papadopoulos et al. [39] implemented content analysis and confirmatory analysis on numerous text sources (i.e., tweets, news, and other social media) to validate the theoretical framework they proposed about explaining sustainability in supply chain. Kim and Kim [40] analyzed news articles and sustainability reports in textile and apparel industry to explore the trends in sustainable supply chain management. As for research addressing supply chain risk management problems, Khan et al. [41] integrated literature review, text mining, and network analysis to develop a framework of strategies and decision making to deal with terrorism-related risks in supply chain management. Su and Chen [42] developed a Twitter Enabled Supplier Status Assessment tool to improve supplier selection process. They applied text mining on Twitter tweets to retrieve supplier related information and then analyze potential risks and uncertainty. Song et al. [43] text-mined the commodity safety regulations of cross-border e-commerce (CBEC), and convert them into risk rules with fuzzy representation. The proposed method was proven to improve the efficiency and accuracy of CBEC commodity risk evaluation.

Sentiment analysis, a subtask in text mining, has also been applied to supply chain management. Sentiment analysis or opinion mining implements text mining and natural language processing (NLP) to discover polarity and recognize emotions; in other words, it helps us understand the tone and opinions of the authors of the textual information [44]. In an application, Chae [45] developed a Twitter analytics framework combining descriptive, content (text mining and sentiment analysis), and network analytics. This study discovered that supply chain related tweets are often posted by supply chain professionals, and some tweets have obvious sentiments about firms' sales performance, and even supply chain disruption. Sentiment analysis can also be used in finance. Schumaker et al. [46] gathered a series of stock prices and the financial news articles that were published at the same time intervals and used sentiments to assist in price direction prediction.

While data analytics and, specifically, text mining or text analytics, have been applied to many supply chain management issues; to the best of our knowledge, the specific research area of comprehensive risk categorization and management in global supply chains has not much benefited from the opportunities these techniques present. One of the reasons for this might be that the global supply chain risk factors have not been well defined; that is, there is a variety of different viewpoints among researchers on this subject. It is crucial to organize previously published concepts of supply chain risk and further define them. Given the characteristics and capability of textual analysis, a text-mining-based risk management framework can bridge the gap of undefined risk factors, risk variation pattern identification, and further supplier selection process. Defining risk factors can be done by using information extraction and text clustering to identify potential risks from the extant literature. Then, risk variation pattern recognition of a region or a company can be accomplished by analyzing sentiments contained in online news articles. This way, the final goal of improving global supply chain risk management could be fulfilled by thoroughly understanding and monitoring the risks.

3. Methodology: global supply chain risk management framework

In this section, a global supply chain risk management framework is proposed and separated into several subsections to illustrate the analysis mechanisms in sufficient details. Figure 1 illustrates the proposed framework. The first step is text data retrieval and preprocessing. The purpose of this study is to analyze texts from extant literature and news articles to assist in global supply chain risk management operations. After data collection and transformation, the second step is to extract and define current potential supply chain risk factors and the risk hierarchy. The last step is risk variation pattern recognition. Resultant risk factors are incorporated with supply chain related online news feeds to discover aspect-wise sentiments and further calculate individual risk levels for each type.

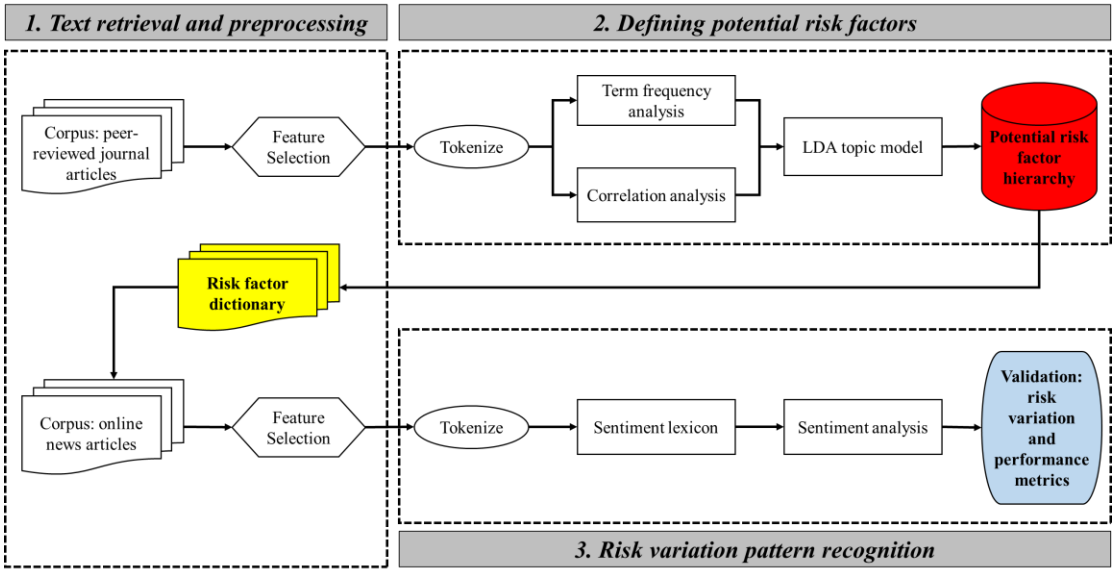

Figure 1. The proposed framework

Throughout this study, R statistical programming language is used to process text data and further analyze and visualize the results. In this research, *tm* and *tidytext* packages [47, 48] are used for the analysis steps of the proposed framework. More details are introduced in the following subsections.

3.1. Defining potential risk factors

3.1.1. Corpus setup: Peered-reviewed journal articles

First a collection of peer-reviewed journal articles related to supply chain risk management was set up as the corpus. The peer-review process coming to journal articles makes them a proper source to discover potential risk factors and their definitions. In this study, Scopus was used to search for articles pertinent to supply chain risk management. Titles, abstracts, and keywords of articles were set as the main search fields. Terms such as supply chain, risk management, evaluation, assessment, and identification were the search grid. In addition, we also limited the results to journal articles that are in English and were published between years 2000 and 2020. As a result, a total of 914 articles were downloaded as the input data of this research.

Although more data tends to have more objective results in the field of data mining, the correlation calculation mechanism in text mining differs from the traditional correlation analysis. In this research, correlation is calculated at document level. In other words, the correlation represents the degree of two terms appearing in the same document. Based on this mechanism, once the number of documents and the total number of words increase, the correlation will be diluted. Therefore, the corpus with

911 documents was further reduced to a corpus with 118 documents, by limiting the title with containing “supply”, “chain”, and “risk”. An even a smaller corpus with only 11 documents (well-cited by peer-reviewed journal articles) was also built in order to obtain more significant correlation results. Note that this smallest corpus was also human-investigated. This more condense corpus was read thoroughly and was considered as the state-of-the-art relevant to supply chain risk management. A summary of this condense corpus can be found in the appendix. We note that across the three corpora the results were stable other than the correlation results, which became smaller with the increasing corpus size.

Before further analysis, feature selection was conducted to reduce the dimensionality of the corpus. In the text mining field, features or attributes represent terms appearing in the corpus, and therefore feature selection can be done by data pre-processing to remove less meaningful terms such as numbers and stop words. Several pre-processing procedures were required to tune the corpus. Since the corpus is in English, all letters were converted into lower case; all numbers and punctuation were also removed. Stop words that are usually used to smooth the sentences were also striped out. All words were then stemmed for simplification. The remaining terms are more condense and meaningful after the pre-processing. Feature selection can also be conducted by considering the sparsity or calculating the term-frequency-inverse-document-frequency (TF-IDF). These two dimensionality reduction approaches serve as alternatives to further improve the analysis results.

3.1.2. Term frequency and correlation analysis

After tuning the corpus, *tm* and *tidytext* packages were used to obtain summary information of the text. First the corpus was tokenized into a per-term-per-document (1-gram) data frame. This format is easy to further calculate term frequency. Bi-gram analysis can provide interesting insights by analyzing the frequency of two consecutive terms. Document term matrix is also a convenient format to calculate term frequency and even correlations between terms. Table 1 illustrates an example of document term matrix from the 11-article corpus. It is a form where rows represent each document and columns represent each term appearing in the corpus. The entries are the number of times that the corresponding terms appear in the corresponding documents. Package *tidytext* provides a lot of useful functions to deal with tidy text format, and it also has a function that is able to convert per-term-per document data frame into document term matrix object, which is a required input format for topic modelling analysis using package *topicmodels* [49].

Article No.	Stemmed term frequency							
	"abil"	"abl"	"abstract"	"accept"	"accomplish"	"accord"	"achiev"	"across"
1	1	1	1	1	1	16	5	1
2	2	1	0	1	0	4	2	10
3	2	1	0	2	0	3	0	1
4	2	0	1	0	0	10	2	2
5	1	0	0	4	0	9	2	0
6	13	3	0	2	0	0	7	12
7	0	0	0	0	7	5	9	25
8	1	1	2	2	0	2	0	6
9	2	1	0	0	0	0	0	0
10	0	0	3	0	0	5	0	0
11	0	1	0	2	1	14	2	2

Table 1. Document-term matrix (partial)

After the tokenizing process and document term matrix conversion, term frequencies were obtained by calculating column sums. From term frequency analysis, information about how frequently a term appears in the corpus or a single document can be obtained. Consequently, whether a term or its synonym is important can be understood. In this research, global supply chain risk management journal articles were analyzed, and therefore, from term frequencies, the important risk factors with their attributes (related terms) can be derived.

Correlation analyses were also employed in this research. Both packages tm and tidytext packages provide functions to calculate correlation. The correlation score ranges from 0 to 1, meaning the degree of how terms are correlated. By interpreting correlations, the relation between two words can be revealed. By investigating correlations, terms having higher correlation with the term "risk" can be considered more related to risk context. This is used to identify risk factors by emphasizing on highly correlated terms. The n-gram analysis is also a useful method to discover relations between terms. The concept of n-gram analysis is from the regular tokenizing process. Initially the corpus was tokenized into a per-term-per-document (1-gram) data frame. N-gram analysis is when not only one term but a fixed number of consecutive terms are tokenized into the tidy data frame. This study employed bi-gram analysis to further discover the connection between terms, especially between the term "risk" and others. An illustration of 1-gram and bi-gram is exhibited in Figure 2.

"Global supply chain risk management"

1-gram	Bi-gram
Global	Global supply
Supply	Supply chain
Chain	Chain risk
Risk	Risk management
management	

Figure 2. An example of 1-gram and bi-gram

3.1.3. LDA topic model

Topic model is a text mining unsupervised clustering technique frequently applied to group unlabeled documents into representative topics. Blei et al. [50] first proposed latent dirichlet allocation (LDA), a flexible generative probabilistic model for text corpora and discrete data. It is a three-level hierarchical Bayesian model, and its application on text mining is to cluster unlabeled documents. The basic concept of topic model is that every document is assumed arising from a mixture of topics, and each topic contains a mixture of terms from the corpora vocabulary. A package in R called topicmodels [49] provide researchers a simple way to derive LDA topic models, and package ldatuning [51] provides functions to tune the parameters, including finding the optimal set of the number of topics. In this research, topic models provide an automatic way to understand the main themes of the documents in the corpus. The use of topic models along with term frequency and correlation analyses resulted in a more comprehensive risk factor categorization dictionary.

3.2. Risk variation pattern recognition

After defining potential global supply chain risk factors, the next step is to recognize risk variation pattern. A variety of global supply chain event information is discussed in business magazines and news articles. Usually, online news articles are updated in very short time intervals. Accordingly, this research focuses on online news articles and conducts sentiment analysis to evaluate global supply chain risk variation pattern.

3.2.1. Corpus setup: Online news articles

Unlike collecting peer-reviewed academic journal articles in Section 3.1., here online news articles were gathered. There are different social media platforms that collect news from various sources. Google News application programming interface (API) was employed to search for news of interest. Google news API was selected due to its several advantageous features: 1) free to use; 2) collects news from diverse sources, including different topics such as finance, business, technology, politics, environment, etc.; 3) it has clear tutorial and instruction documentation. Thus, this study selected Google News API as the source of online news articles.

The defined risk types and the underlying risk factors structured from the former analysis served as a risk factor dictionary. This dictionary helps filter out news about specific topics to further analyze supply chain risks for a particular company, industry, or region. Specifically, the news articles in the hardware technology industry sector published in 2018 were collected into the corpus for this research.

The feature selection process (data pre-processing) was implemented as described above. Numbers, stop words, and punctuations were removed. The relationship and trend between article sentiments over time are essential for capturing risk variation pattern. Therefore, every article in the corpus was indexed by their publication date also. This way sentiments can be seen over a time horizon and connections can be made to critical decisions.

3.2.2. Sentiment analysis

Sentiment analysis is a method for calculating polarity or discover emotions from words [44]. It can be implemented at either corpus level, document level, section level, or even sentence level. This study incorporates lexicon-based sentiment analysis to capture supply chain risk variation patterns, using predefined sentiment lexicons [52].

A sentiment lexicon comprising a list of words, assigned with different sentiments, is required for sentiment analysis. In R package tidytext, there are several sentiment lexicons, namely nrc, bing, AFINN, and loughran. The nrc lexicon categorizes its sentiment word list as "positive", "negative", "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", or "trust"; the bing lexicon has binary values of "positive" or "negative"; AFINN uses numerical scores to reflect sentiments, ranging from -5 to 5, representing "positive" and "negative"; and loughran has categorical values of "litigious", "uncertainty", "constraining", and "superfluous". These lexicons were implemented individually on the news article corpus to generate sentiment results. In the end, overall sentiments of the corpus and daily polarities over time were investigated.

3.2.3. Validation

The sentiment analysis results require further validation to ensure that sentiments triggered by news articles are relevant to the proposed risk factors. However, for different global supply chain risk types, researchers and practitioners may apply various other performance metrics for evaluation. In order to highlight the synergy of the proposed framework, the scope of the risk was narrowed down to financial and operational aspects, and correspondingly, stock price was used for validation. Stock price is volatile and can be affected by a variety of factors. Stock price can also be seen as investors' expectation of a company's capability of operation and profitability. When something risky to the corporation or negative events happen, it's likely that it would be reported in a news article. Indeed, given the plethora of investment companies and news organizations, this reporting is expected. Therefore, by using the risk factor term dictionary developed from our risk categorization hierarchy, news related to specific risk type can be filtered out, and a particular company can be selected. By comparing the sentiment polarity of those news articles and stock price of the company within the same time interval, the risk variation pattern can be recognized and validated. Of course, the underlying assumption is that investors are informed about the related market events with implications on the company and that realized stock price is a reflection of investor beliefs.

4. Results and discussion

4.1. Potential risk factors: Global supply chain risk hierarchy

In this section, results of the Defining Potential Risk Factors phase were illustrated and discussed. Term frequency analysis, correlation analysis, and LDA topic modelling were conducted on three corpora, the 11-article-human-investigated one (corpus A), the 118-article-machine-selected one (corpus B), and the 911-article-initially-machine-selected one (Corpus C). The reduction of the corpus size from Corpus C to B and then to A was implemented in order to gauge the validity of the results. Based on the mechanism of document level correlation, the more documents and words in a corpus, the lower the correlation would be. However, if the corpora were appropriately chosen, the identified risk factors, by and large, would remain the same. All three corpora were used for the entire analysis. A summary of the subject corpora is shown in Table 2.

Dimension	Corpus A	Corpus B	Corpus C
Number of documents	11	118	911
Number of terms	154,429	1,341,628	10,368,960
Number of terms (after pre-processing)	57,434	531,310	5,008,339

Unique stemmed terms	4,927	21,205	128,782
Unique stemmed terms (after TF-IDF reduction)	2,205	10,351	53,655

Table 2. Summary of the corpora

All corpora were pre-processed by removing numbers, punctuations, stop words, and terms that appear regularly in journal articles such as “introduction” and “references”. Regular expressions were used to allow the corpora only contain the main body of the articles, filtering out the reference list and other irrelevant terms. Only meaningful terms were retained. Document term matrices were generated for further analysis. As shown in Table 2, numbers of terms in all corpora reduced significantly after pre-processing. Unique stemmed terms were used for term frequency and correlation analysis, and TF-IDF reduction was conducted on these stemmed terms for topic modelling.

Term frequency were conducted to all corpora. Figure 3, 4, and 5 demonstrate the distribution of term frequency and word cloud visualization. An implication from these figures is that although they are independent (i.e., there are no identical articles within the corpora), there are many frequent terms that are similar. This implies that the condensed corpus A is a valid representative of the supply chain risk management research field, even though it contains very few documents. Therefore, the analysis results from corpus A was considered as the primary structure for defining potential risk factors complemented by the results from corpus B and C.

Table 3 presents the identical and different terms of high frequency (partial) to further explain Figures 3, 4, and 5. The identical terms are reasonable and include “risk”, “suppli”, etc. They result from the scope of the corpus development focusing on supply chain risk management. Different high frequency terms explain that the corpus A mainly contains articles that applied literature review methods to categorize risk factors, and the corpus B set their focus more on risk assessment and mitigation, especially in specific business sectors.

Identical terms	“risk”, “suppli”, “chain”, “manag”, “product”, “supplier”, “model”, “process”, “cost”, “intern”, “strategi”, “factor”,“decis”, “oper”, “analysi”, “approach”,
Different terms	“global”, “sourc”, “company”, “review”, “logist”, “inform”, “industri”, “disrupt”, “level”, “system”, “mitig”

Table 3. Comparison of terms of high frequency among corpus A, B, and C (partial)

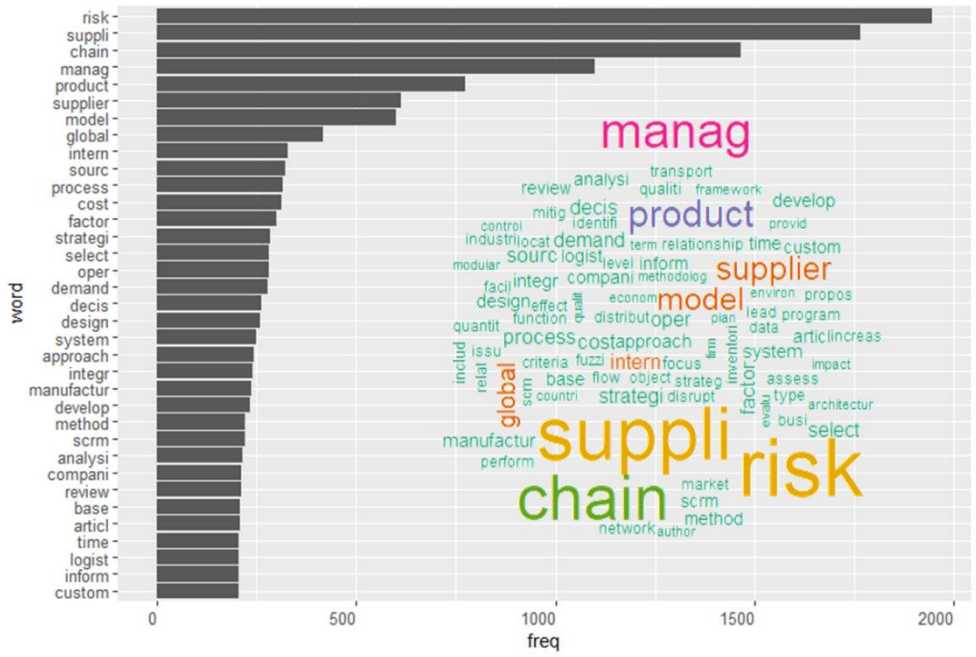

Figure 3. Term frequency of corpus A

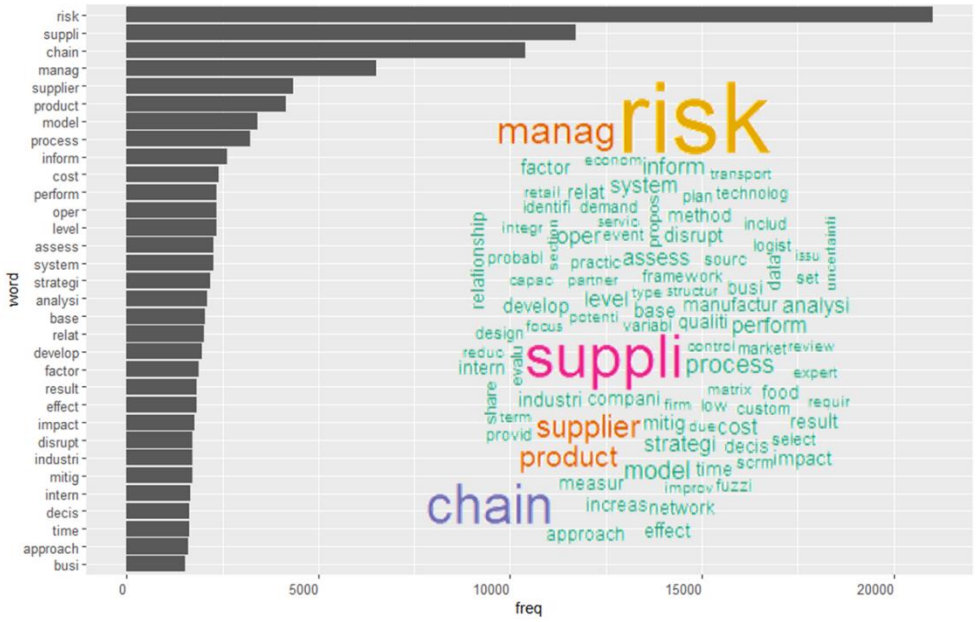

Figure 4. Term frequency of corpus B

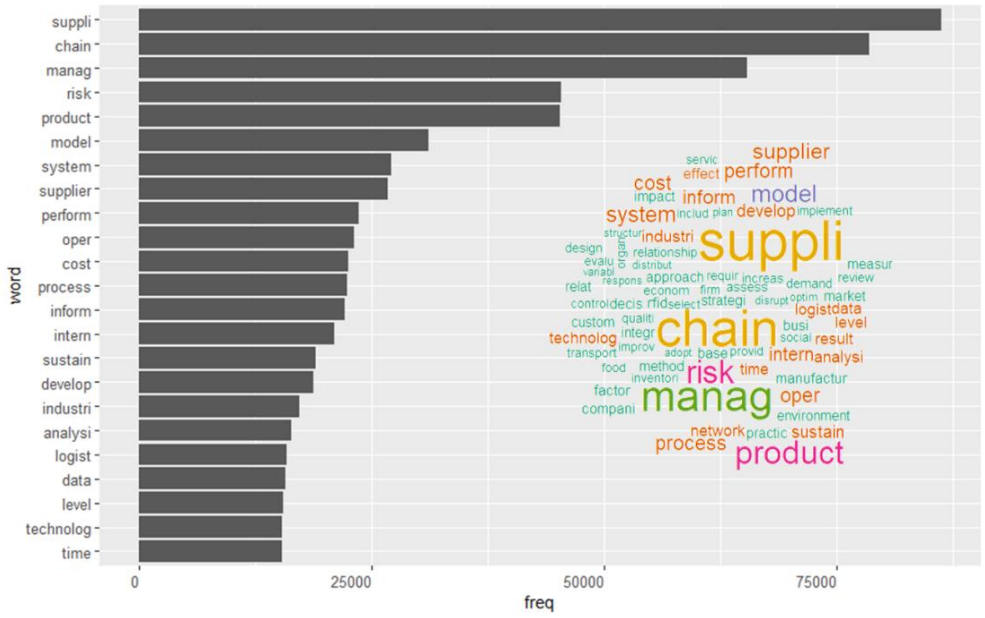

Figure 5. Term frequency of corpus C

In order to discover potential risk factors, correlations between terms (specifically with the term "risk") were analyzed. The terms having correlations greater than 0.6 were examined and are listed in Table 4. All 1,030 entries were investigated and 84 terms related to supply chain management were screened out.

Stemmed / original words	Correlation	Stemmed / original words	Correlation
"terror" / terrorism and terrorist	0.93	"micro" / micro	0.83
"demand" / demand	0.92	"nuclear" / nuclear	0.83
"weather" / weather	0.92	"obsolesc" / obsolescence	0.83
"extern" / external	0.90	"oper" / operation	0.83
"breakdown" / breakdown	0.88	"organisa" / organization	0.83
"replenish" / replenishment	0.88	"overstock" / overstock	0.83
"labor"/ labor	0.87	"profit" / profit and profitable	0.83
"utilis" / utilization	0.87	"schedul" / schedule	0.83
"macro"/ macro	0.86	"technolo" / technology	0.83
"account"/ account and accounting	0.85	"technic" / technical	0.83
"disast" / disaster	0.85	"telecommun" / telecommunication	0.83
"disrupt"/ disruption	0.85	"turbul" / turbulence	0.83
"retail"/ retailing and retailer	0.85	"budget" / budget	0.82
"social"/ social	0.85	"yield" / yield	0.81
"variat" / variation	0.85	"insur" / insurance	0.81
"war" / war and warfare	0.85	"competitor" / competitor	0.79
"chain" / chain	0.84	"earthquake" / earthquake	0.78
"maritim" / maritime	0.84	"bankruptcy" / bankruptcy	0.77
"strike"/ strike	0.84	"natur" / nature and natural	0.77
"suppli" / supply and supplier	0.83	"efficienc" / efficiency	0.76
"financ" / finance and financial	0.83	"hurrican" / hurricane	0.76
"flood" / flood	0.83	"instrument" / instrument	0.76
"fire" / fire	0.83	"qualit" / quality	0.76
"flow" flow	0.83	"scarc" / scarce	0.,76
"fluctuat" / fluctuate and fluctuation	0.83	"short" / short and shortage	0.76

“abnorm” / abnormal and abnormality	0.83	“substitut” / substitute	0.75
“bottleneck” / bottleneck	0.83	“credit” / credit	0.75
“bullwhip” / bullwhip	0.83	“intern” / internal	0.74
“capac” / capacity	0.83	“assest” / asset	0.73
“catastroph” / catastrophe	0.83	“manag” / manage and management	0.73
“deficient” / deficient	0.83	“econom” / economic	0.72
“disassembl” /disassembly	0.83	“network” / network	0.72
“disclosur” / disclosure	0.83	“uncertainti” / uncertainty	0.72
“downturn” / downturn	0.83	“capit” / capital	0.71
“ecosystem” / ecosystem	0.83	“price” / price	0.70
“efficaci” / efficacy	0.83	“transpar” / transparent	0.67
“employe” / employee and employer	0.83	“transport” / transportation	0.67
“horizont” / horizontal	0.83	“safeti” / safety	0.66
“inflat” / inflation	0.83	“transact” / transaction	0.65
“inform” / information	0.83	“dissatisfact” / dissatisfaction	0.64
“infrastruc” / infrastructure	0.83	“environment” / environment	0.64
“loan” / loan	0.83	“tsunami” / tsunami	0.61

Table 4. Terms highly correlated with “risk” (correlation threshold = 0.6)

Take “terror” and “war” (having 0.9-ish correlation with “risk”) as examples, a global supply chain has stakeholders throughout the world, and devastating events such as terrorist attacks and wars would significantly impact the local suppliers at the event regions. Consequently, this will have tremendous influence on supply chain network, from raw material supply, economic breakdown, to destruction of firms’ infrastructure.

Environmental changes and climate variations are also crucial to global supply chain networks. Terms such as “weather”, “flood”, “earthquake”, and “tsunami” can be categorized as environmental risks, and such natural disasters would, for instance, damage suppliers’ inventory, raw material and manufacturing facilities, causing operational challenges and greatly affecting the supply chain. The correlation results above can be seen as inputs to yield a comprehensive dictionary for us to refer to while defining supply chain related risk factor attributes.

Although correlation analysis provided meaningful results for constructing and defining potential supply chain risk factors, the stemming feature selection process of the corpora and multiple meanings of words might lead to different interpretations. Most of the time when introducing risk factors, phrases with two words such as “supply risk” or “inventory risk” are used. In order to take this issue into consideration, a bi-gram analysis was applied to discover how often a word appears consecutively with “risk”.

Figure 6 demonstrates the frequent terms that appear right after “risk”, and Figure 7 displays the frequent terms appearing right in front of “risk”. As shown in Figure 6, the phrases “risk type” and “risk mitig” appear more than 50 times and 20 times,

respectively, indicating the corpus highlights discussion of risk types and risk mitigation. Other terms that come after risk also provided information on the subjects discussed in the articles while talking about "risk". Moreover, terms that come directly before risk give us more intuitive information on the potential risk factors. In Figure 7, there are several frequent terms that come before "risk" that demonstrate interesting connections with risk factors, such as "sourc" (e.g., source), "suppli" (e.g., supply), "chain", "financi" (e.g., financial), "demand", "manufactur" (e.g., manufacture), "environment", "infrastruct" (e.g., infrastructure) and so on. Figure 8 visualizes the bi-gram network (only shows the bi-grams that occur more than 20 times) for easy detection of relationships between terms.

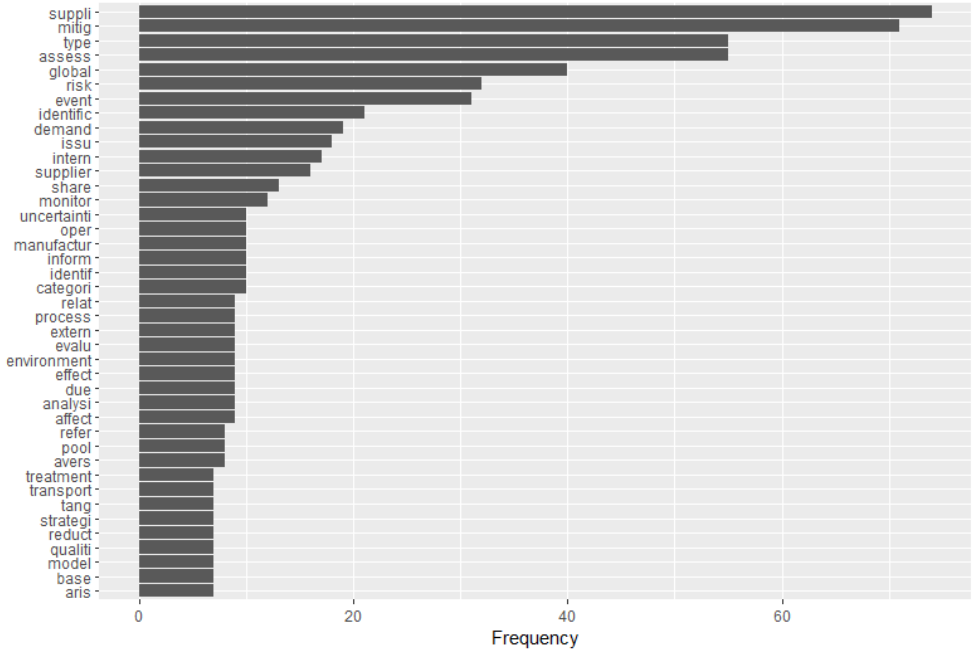

Figure 6. Frequent terms that come after "risk"

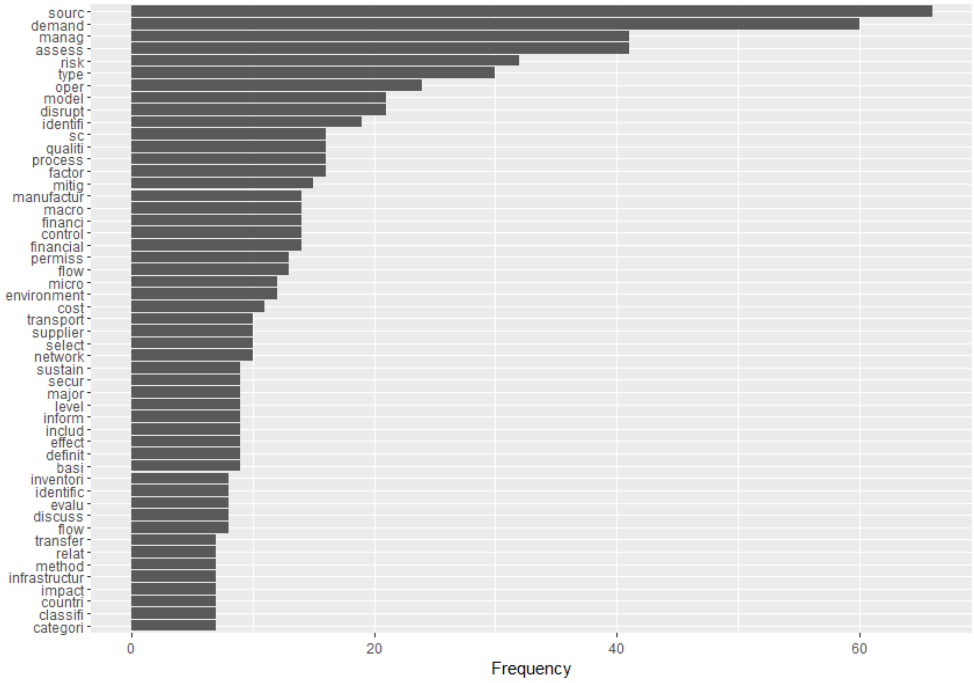

Figure 7. Frequent terms that come before “risk”

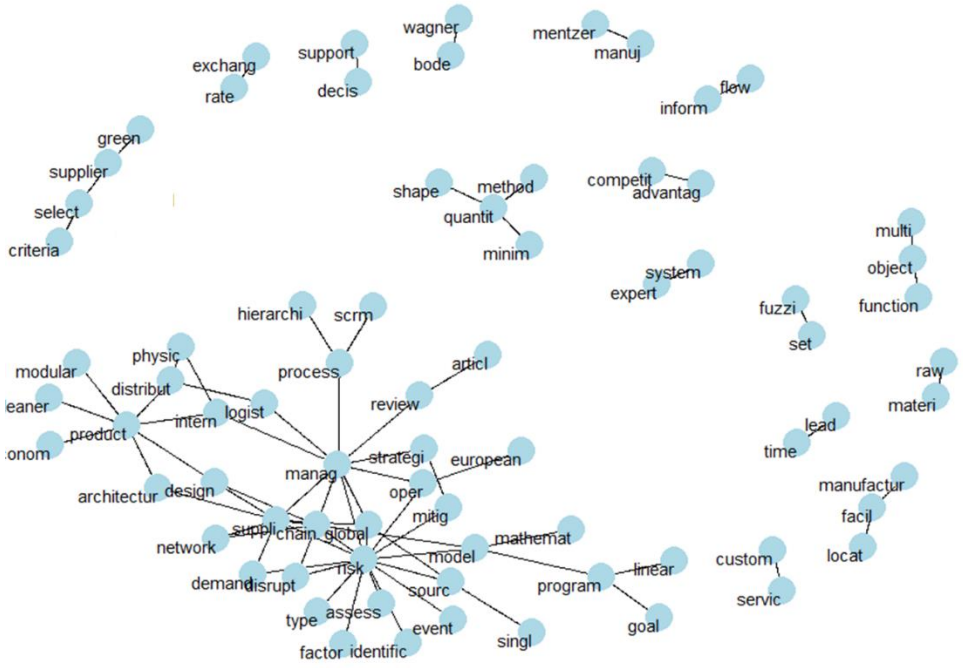

Figure 8. Bi-gram network

Since there is a great difference in number of documents among different corpus, an unsupervised learning algorithm, can assist in understanding the main themes of each document and grouping them into several topics. Corpus A has only 11 articles, and is a condensed collection; it therefore provides less information about topics when topic modelling is used. Corpus C has 911 articles, but applying topic modelling on

large corpus would lead to low interpretability. When the optimal number of topic is too large, decision makers would have a hard time interpreting the topic clustering results. Accordingly, LDA topic modelling was conducted on corpus B to generate topic clustering to have a more inclusive understanding of each individual document.

LDA topic modelling requires a fixed prior-set number of topics, similar to other clustering approaches. There isn't a definite way to select the number of topics for LDA model. However, there are several metrics such as KL-divergence, density-based inherent connection, and Bayesian statistics, that can help determine the optimal number of topics, and they are built in package ldatuning [53, 54, 55, 56]. Package ldatuning [51] was used to determine the optimal number of topics. The result of the tuning process shows that clustering performance is consistently optimal when the number of topics is from 20 to 80. In this research, topic modelling served as a complimentary tool for discovering the mixture of topics of the corpus, and assuring there was no important risk-factor-related terms being neglected from the former analysis. Thus, the number of topics was set to 20. This setting also lead to better readability compared to a large number of topics.

Table 5 presents 20 of the topics generated from corpus B, and five representative terms are listed under each topic. TF-IDF was used to reduce dimensionality. The original document term matrix of corpus B contained 21,205 unique terms, and after TF-IDF screening process the matrix was reduced to 10,351 terms. Terms such as "risk", "supply" and "chain" appear in nearly all documents, leading to having low TF-IDF, and therefore were removed. These words are relatively less important for topic modelling since they appear in most of the documents in the corpora.

The topic model revealed the main themes of the documents in corpus B. These topics indicate that the current supply chain risk management field has been focusing on specific industry sectors' risk management; this can be explained by the terms such as "steel", "food", "maritime" and "hardwar". From the results, it is also clear that sustainability and green supply chain are among the high interest areas. It can also be observed that corporate management, strategy, and relationships with stakeholders in their supply chain are also essential given the terms "tactic", "organis", "stakehold", "trust", and "partnership". A variety of decision support methodologies (Delphi-method, analytic hierarchy process, failure mode and effects analysis, fuzzy theory and Bayesian method) appear in the topics. This implies that researchers and practitioners have focused on aggregation methods for risk quantification and uncertainty inherent in risk factors.

LDA Topic model				
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
“stakehold”	“steel”	“pig”	“green”	“delphi”
“sustain”	“flow”	“barrier”	“fuzzi”	“attack”
“trust”	“china”	“crime”	“fmea”	“metric”
“scm”	“pharmaceut”	“humanitarian”	“partnership”	“stakehold”
“firm”	“turbul”	“organis”	“tactic”	“hardwar”
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
“fuzzi”	“countri”	“food”	“contract”	“water”
“food”	“farmer”	“resili”	“retail”	“financial”
“oil”	“wheat”	“agri”	“expertis”	“motor”
“maritim”	“corpor”	“entropi”	“payment”	“crisi”
“ahp”	“bin”	“perish”	“profit”	“food”
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
“scrm”	“rfid”	“trigger”	“food”	“resili”
“fuzzi”	“retail”	“resili”	“node”	“legitimaci”
“firm”	“contract”	“oem”	“retail”	“plant”
“influenc”	“wholesal”	“profit”	“tier”	“wind”
“matrix”	“profit”	“accuraci”	“bayesian”	“sustain”
Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
“firm”	“ahp”	“food”	“terror”	“food”
“flexible”	“criteri”	“hazard”	“resili”	“dairi”
“rerout”	“hydroge”	“climat”	“biomass”	“hazard”
“vehicl”	“cluster”	“bioenergi”	“algorithm”	“green”
“electr”	“pairwis”	“energi”	“bender”	“ahp”

Table 5. Topic model (k = 20)

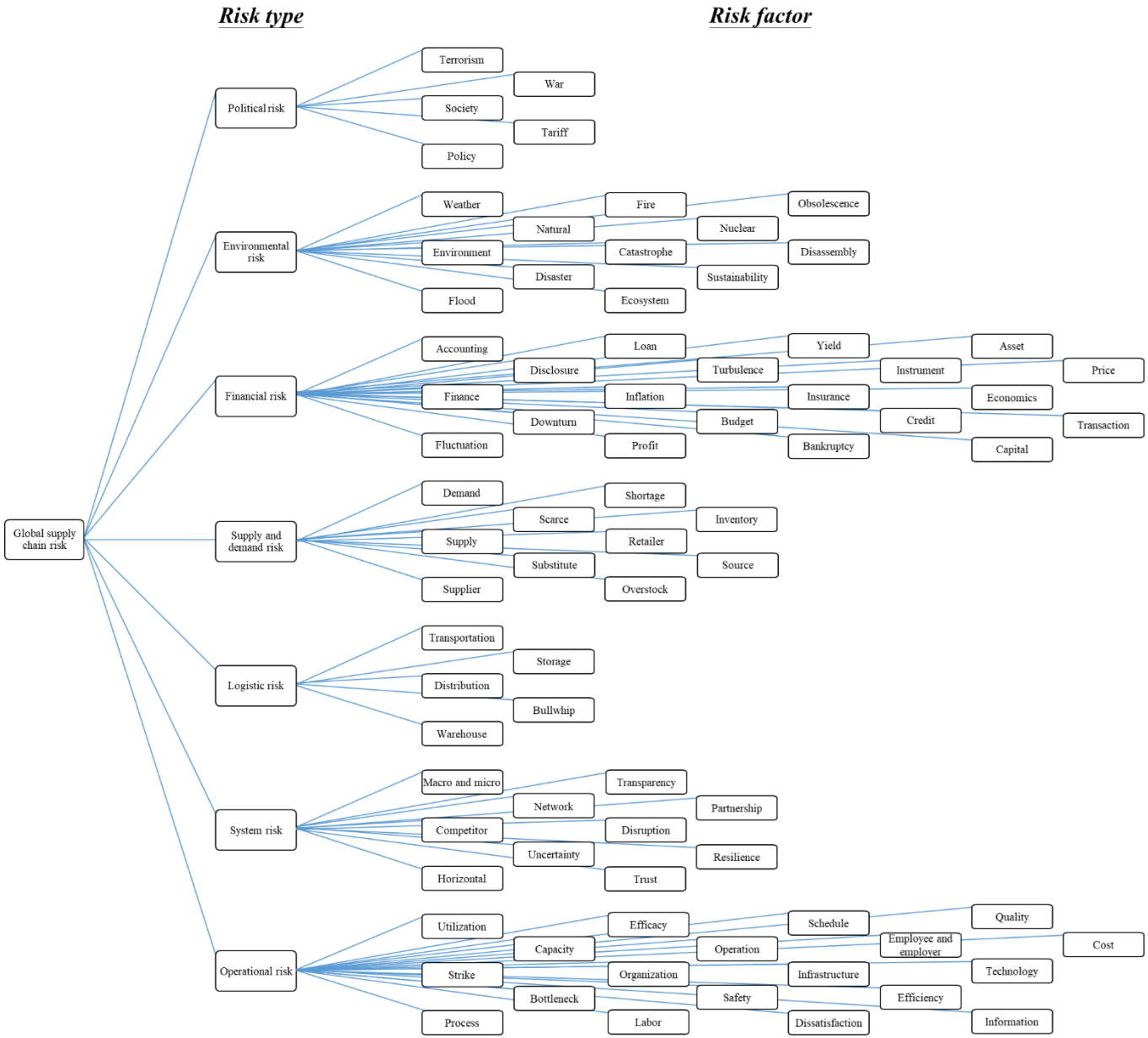

Figure 9. Proposed risk type and risk factors

Based on the results from term frequency and correlation analysis, bi-gram analysis, and LDA topic modelling, global supply chain risk was categorized into seven types and each type contains several potential risk factors derived from critical stemmed terms, as shown in Figure 9.

In the proposed hierarchical categorization, global supply chain risk is decomposed into Political, Environmental, Financial, Supply and demand, Logistic, System, and Operational aspects. Each aspect has its representative terms as its risk factors. For example, while events related to terrorism or tariff happen in a specific time and region, they could potentially impact the whole supply chain network and further affect the performance of relevant stakeholders.

One other example from the categorization is that environmental risk would increase when a region is struck by a natural disaster like an earthquake or a hurricane. The frequency of a particular natural disaster in a region would also affect the potential environmental risk. If a region is used to have earthquakes or hurricanes, the local manufacturing companies would have high environmental risks, and therefore they would probably take actions to mitigate the possible damage that might be imposed by the disasters. This implies that even though the global supply chain risk types and factors are categorized by a hierarchical structure in this study, these risks types and factors may not be independent. In other words, several risk factors or risky events would have chain reactions impacting the entire or part of a global supply chain, and the fundamental information and composition of a supply chain (industry types, product types, involved regions, etc.) would be consequently crucial in order to discover the connection among those risk factors.

Although the proposed global supply chain risk hierarchical categorization did not reveal inner-connections among the risk elements, it is fundamentally important for enterprises to understand different risk types and their underlying factors. The proposed risk management framework reorganized the state-of-the-art research articles of supply chain risk management by text mining techniques, and developed a comprehensive categorization for potential risk types and factors. This risk categorization can therefore become guidance to companies for their supply chain risk management operations.

In the following section, this categorization dictionary was used as a classifier for filtering out specific online news articles from Google News related to particular supply chain risks, and then sentiment analysis was conducted for capturing risk variation.

4.2. Risk variation pattern recognition: Sentiment analysis on Google News

This section demonstrates how risk variation pattern recognition phase in the proposed framework can be implemented to improve global supply chain management.

First, Google News API was used on the collected news articles. News articles related to hardware technology industry sector were filtered out, and the search parameters were set to obtain financial and operational risk related documents. Risk factor terms such as finance, profit, revenue, technology, capacity, and cost were the keywords for the search. Moreover, the subject of articles was narrowed down to focus on only one company (Yageo® Corp.) and limited to the publication year of 2018. This setting filtered out 69 news articles in total. Through a brief scanning of articles (human-investigation), a total of 10 identical (Google News API have access to different sources of news so it might have repeated articles) or irrelevant articles were eliminated. A total of 59 news article about financial and operational aspects were then set as the study corpus. An example article is shown below. This example provides a brief background introduction of the company:

"A Taiwanese manufacturer of components used in everything from smartphones to computers has turned into the world hottest stock amid surging demand for its products. Yageo Corp. has jumped more than 800 percent since the start of last year, the most among stocks on MSCI Inc. global gauge, to add $8.1 billion of value. The company, which makes capacitors and resistors that are essential for regulating electrical flows within electronic products, boosted profit more than five-fold from a year earlier in the first quarter as competitors cut output. Chinese production of such components has been reduced as the government moved to curb pollution, while leading Japanese manufacturers have shifted capacity into other products, creating a shortage that has benefited Yageo, according to Allan Lin, assistant vice president at Concord Securities Co. The result is an "astonishing" increase in earnings per share, Lin said. Analysts continue to be bullish on the company, which supplies companies including Apple Inc., Intel Corp., and Sony Corp. There are 12 buys, zero holds and just one sell on the stock. The firm reported a 99 percent increase in sales in April from a year earlier. "The gain in our stock is determined by the market based on our earnings results," Yageo spokesperson Sandy Chang said by phone. Chairman Pierre Chen has a 7.7 percent stake, according to the latest data compiled by Bloomberg. --- With assistance by Sofia Horta E Costa, and Young-Sam Cho

After collection of articles, the corpus was tokenized into a tidy data frame, and all numbers, punctuations, and stop words were removed. It initially had 31,273 words; after data preprocessing, only 15,954 words (which 3,345 of them are unique words) were left for sentiment analysis.

Package *tidytext* exists 4 built-in lexicons, *nrc, bing, AFINN,* and *loughran*. Table 6 provides the basic information of these lexicons. *AFINN* uses numeric values from -5 to 5 to reflect sentiments, *bing* only uses "positive" and negative, *loughran* specify sentiments in "constraining", "litigious", superfluous, and "uncertainty", and *nrc* is categorized in not only "positive" and "negative" but also emotions like "anger" and "disgust". Besides categories, *nrc* contains more words than other three lexicons; followed by *bing*; and the *AFINN* has the least.

Table 7 illustrates the match ratio between the individual lexicons and the news article corpus. Generally, the match ratio is low. It's because not every word has sentiment and there is no perfect general-purpose lexicon with suitable sentiment words for a specific domain. To further improve match ratio, a specific domain lexicon can be generated by sentiment classification. Since *nrc* and *bing* have better matching performance, they were used for sentiment analysis.

Lexicon	**Words in lexicon**	**"anger"**	**"anticipation"**	**"constraining"**	**"disgust"**
AFINN	2,476	NA	NA	NA	NA
bing	6,785	NA	NA	NA	NA
loughran	3,916	NA	NA	184	NA
nrc	6,468	1,247	839	NA	1,058
	"fear"	**"joy"**	**"litigious"**	**"negative"**	**"positive"**
AFINN	NA	NA	NA	1,597	879
bing	NA	NA	NA	4,782	2,006
loughran	NA	NA	903	2,355	354
nrc	1,476	689	NA	3,324	2,312
	"sadness"	**"superfluous"**	**"surprise"**	**"trust"**	**"uncertainty"**
AFINN	NA	NA	NA	NA	NA
bing	NA	NA	NA	NA	NA
loughran	NA	56	NA	NA	297
nrc	1,191	NA	534	1,231	NA

Table 6. Word counts and categories of lexicons

Lexicon	**Matched words**	**Words in articles**	**Match ratio**
AFINN	267	3345	0.08
bing	370		0.11
loughran	327		0.10
nrc	589		0.18

Table7. Match ratio of lexicons and the corpus

By joining the lexicons (*nrc* and *bing*) to the tidy data frame of the news article corpus, words matched with the sentiment word list were counted, and sentiments were calculated for each category by summing them up. The corpus level sentiments applying *nrc* were calculated as shown in Figure 10. The corpus level sentiments seem to be very positive with over 1,500 positive words, twice as many as the negative words. Moreover, presence of emotions such as "trust" and "anticipation" show that a great proportion of the news articles convey a promising perception of Yageo® Corp. on

financial and operational aspects as these were chosen as the scope of the corpus. This result implies that Yageo® Corp. had less stress on these two aspects in terms of sentiments in 2018, and this could be interpreted as the company had less risk in terms of finance and operations.

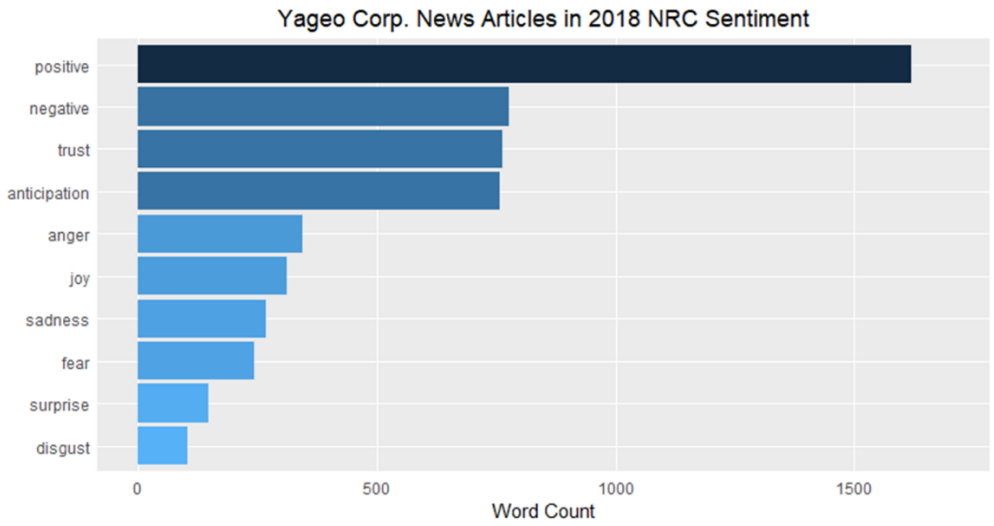

Figure 10. Corpus level sentiments (*nrc*)

Although in *nrc* lexicon, the corpus level sentiment appears to be extremely positive, the positive word count is only slightly greater than the negative words applying *bing* lexicon (Figure 11). A possible reason is that lexicon *bing* actually has more negative words than *nrc* does. Both lexicons have similar total numbers of words, but bing has 4,782 negative words and *nrc* only has 3,324. On the other hand, *bing* also has less positive words (2,006), than *nrc* does (2,312). This causes the differences between these two corpus level sentiment results.

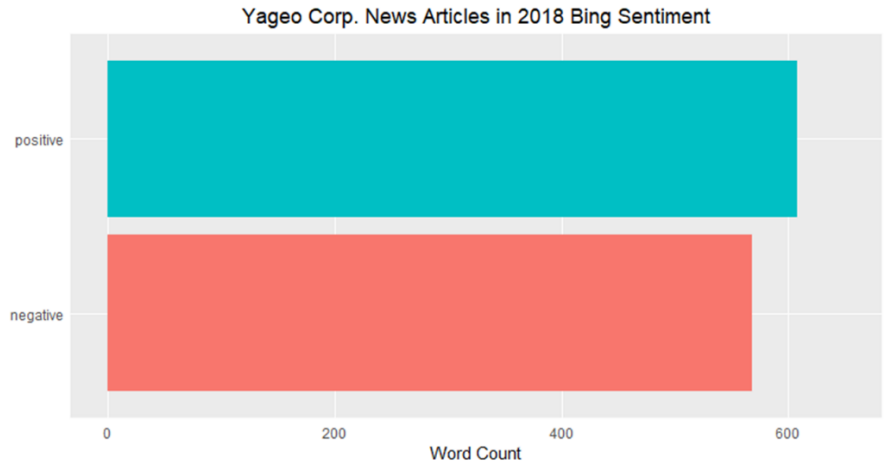

Figure 11. Corpus level sentiments (*bing*)

Other than overall sentiments, polarity and percent positive over time can also

provide insightful information. In this analysis, lexicon *bing* was used because polarity is only related to positive and negative sentiments. The objective of this analysis was to evaluate sentiments on the temporal horizon. Therefore, the articles were ungrouped and regrouped in terms of tokens by date. From the 59 articles corpus, there were only 50 unique dates. The sum of the sentiment word count was then obtained. Polarity for each date was calculated using date-wise positive count minus date-wise negative count. Each date's percent positive was obtained by taking the percentage of date-wise positive count over the sum of date-wise positive and negative counts. Figure 12 illustrates the polarity and percent positive on the date level.

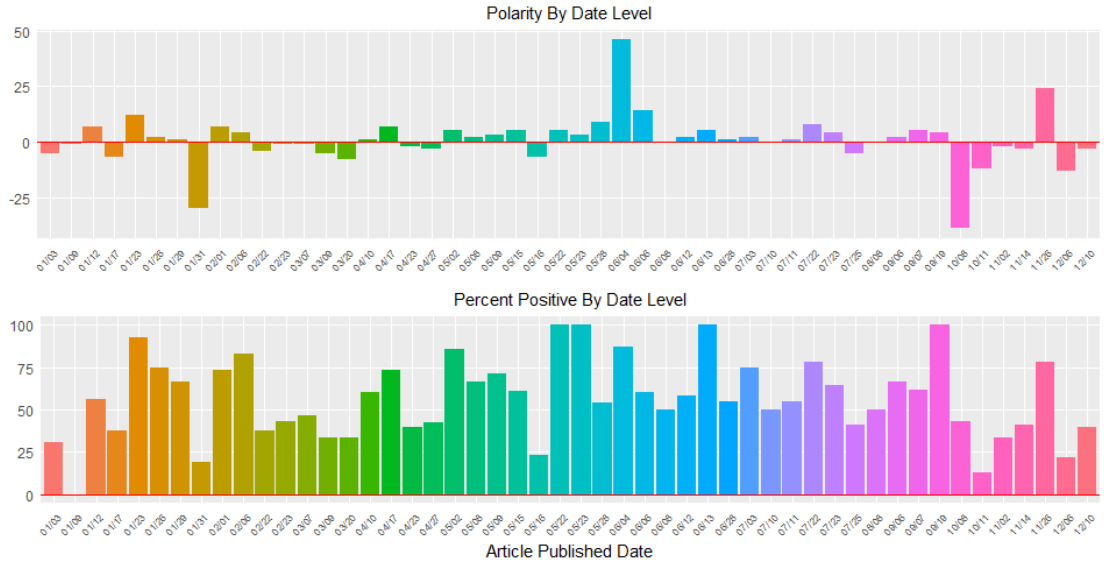

Figure 12. Polarity and percent positive over time using *bing* (daily)

This figure shows that throughout the 50 dates in 2018, the polarity was gently varying. There were only several severe differences in January (negative), May (positive), October (negative), and November (positive). In May, the polarity had the largest variation moving toward positive. However, because the number of articles published and collected weren't the same for each date, there was a slight bias in the figures. For example, in percent positive figure, 5/22 and 5/23 had high percent positive value, but they had relatively low polarity (still positive). In these two dates, they are lower number of articles, and even though the percentage of positive is high, the difference between positive and negative could still be relatively low compared with other dates having more articles published. Despite this issue, this result still implies that authors of these articles were writing more positive than negative news in 2018.

In order to validate sentiment results and their relationship with the corresponding risk factors, historical stock prices of Yageo® Corp. in 2018 were plotted for comparison. Since the corpus is mainly about financial and operational aspects, collected by using

the proposed risk types and factors, stock prices may be an appropriate indicator. As shown in Figure 13, Yageo® Corp. had a significant jump in the middle of 2018, and this was actually its highest point throughout the past few years. The variation and trend in this figure also match the sentiment results. High points and low points of stock prices are clearly similar to the pattern of polarity and percent positive results. In addition, the sentiment polarity and stock price in this time interval are correlated; a 0.30 correlation score with p-value of 0.04 demonstrates a significant moderate positive correlation. This significant correlation indicates that sentiments could reflect Yageo® Corp. stock price trend. Although the stock prices are influenced by numerous factors, negative events or news about potential risks published on news articles may dominate investors' confidence and expectation level of the target company's performance. Therefore, stock price could become one of the evaluation approaches to capture its financial and operational risk variation patterns.

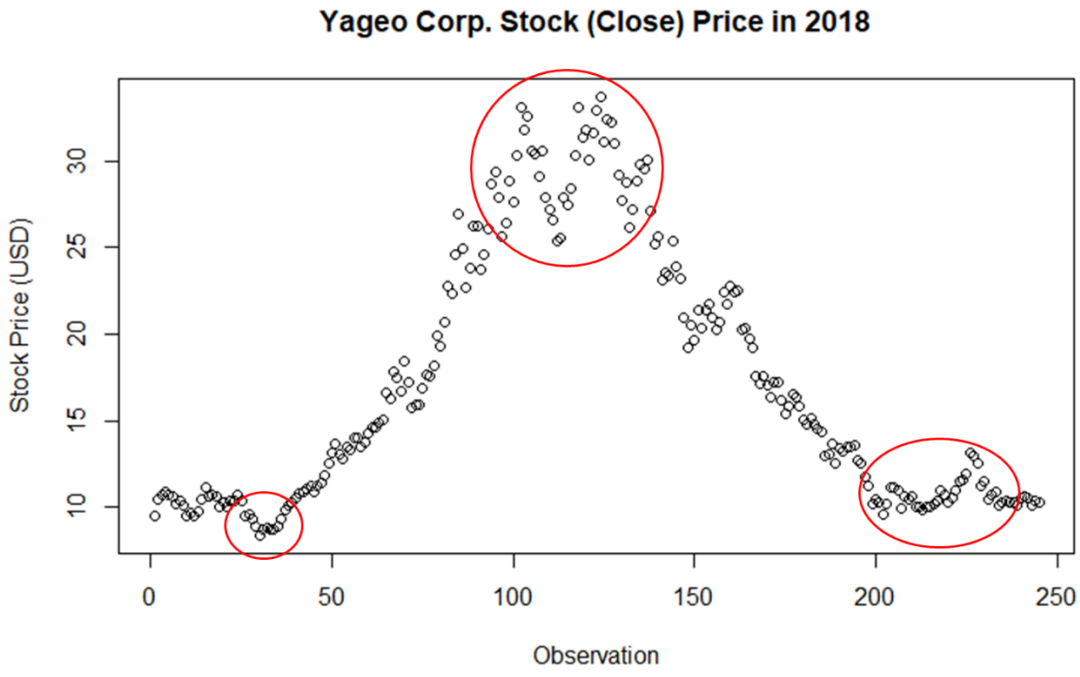

Figure 13. Yageo® Corp. stock price in 2018

4.3. Discussion

This study demonstrates the application of the proposed global supply chain risk categorization and the implementation of sentiment analysis on risk variation pattern recognition. After implementing frequency and correlation analysis and topic modeling on a total of 911 journal articles related to global supply chain risk, a holistic risk categorization of seven risk types and 81 risk factor terms was developed. This

categorization informs industry practitioners and academic researchers about the wide spectrum of risk factors relevant to global supply chains. Under the globalized economic environment, some of these risk factors are extremely region-sensitive, such as natural disasters, government policies, tax regulations, currency exchange rates, etc. Therefore, it is essential that firms manage their operations and develop risk mitigation strategies to account for these uncertainties. The proposed framework can assist firms in automatically and agilely developing a customized and comprehensive risk hierarchy; further, they can target the critical ones (can be judged by industry specialists or decision makers) to manage risk.

The risk factor terms are also a dictionary of global supply chain risk, which can be used to filter news articles that relate to different risk types. This study applied the risk hierarchy to collect 59 news articles from Google News API that discussed Yageo® Corp.'s financial and operational aspects of risk in 2018. The result of the sentiment analysis shows that the sentiment polarity is positively correlated with the company's stock price throughout the year. Stock price is an indicator of the investors' expectation of the company's performance, and when risk events having negative performance consequences occur, it would be affected accordingly. The correlation between sentiment polarity and stock price implies that sentiments deduced from the news articles could signal supply chain risk. Practitioners such as supply chain analysts and managers could be benefit from such signals as they make short and long term decisions. A more accurate risk prediction model could also be developed based on the results of the sentiment analysis.

The presented framework fills several research gaps. For example, Rangel et al. [57] proposed a supply chain risk classification and sorted the risks into five management processes in the supply chain (plan, source, make, deliver and return); however, they applied qualitative literature review on 16 different risk classifications developed in previous research. By contrast, the proposed systematic artificial-intelligent-based framework can generate results more efficiently since it has the power to analyze a larger dataset in a less-time-consuming manner. The results from a data-driven method would also have less human bias. In cases, where data is mostly human judgement-based (e.g., survey data such as presented by [58]), concerns in relation to potential bias may arise. In addition, several studies applied data-driven approaches or analyzing big data in supply chain management or risk analysis field, but they either focus only on intrinsic processes of supply chain, specific industry, or evaluation based on internal numeric data collected from case companies [29, 59]. In contrast, the proposed framework developed a holistic risk categorization supported by a relatively large dataset covering extant literature from year 2000 to 2019, which include multiple

viewpoints on global supply chain risks and a variety of industry sectors. Finally, although there are a few supply chain risk studies that incorporated text mining (e.g., Khan et al. [41] focused on terrorism for its effect on supply chain risk, and Wu et al. [33] tackled sustainability), to the best of our knowledge, there has never been an attempt to develop a data-driven comprehensive supply chain risk framework. Moreover, there is no previous research incorporating sentiment analysis of news articles to investigate risk variation patterns. Global supply chain risk management has become an important research focus, and by filling the identified research gaps, this study provides a holistic supply chain risk framework for companies to develop strategies to manage risk.

There are several limitations in the study which could be addressed in the future research. First, the proposed risk categorization didn't discover the dependency between risk factor types. Risk factors of one type could also have influences on other types. Second, in risk variation pattern recognition phase, the subject was narrowed down to one specific company in hardware technology sector for analysis. This is for the purpose of demonstrating the implementation of the proposed framework. Global supply chain is a system with numerous stakeholders, and hence the risks should also be interconnected. The company's upstream and downstream stakeholders should also be taken into consideration in terms of risk variation capturing. Second, there were only 59 news articles in 2018 discussing about Yageo® Corp. financial and operational aspects. Additional sources and more risk aspects could be added to the analysis to generate more comprehensive results. Finally, stock price was used to validate the efficacy of sentiment analysis. However, for different aspects of risks and different type of focal companies, additional metrics should be considered to evaluate the analysis performance.

5. Conclusion

This research aimed at utilizing text data from extant literature and online news articles to address global supply chain risk management. Significantly, a global supply chain risk management framework was proposed. The framework has three essential phases: First, two main corpora were set up of peer-reviewed journal articles and online news articles, respectively. Feature selection (data pre-processing) was conducted to remove noise terms and reduce dimensionality. Second, research journal articles related to global supply chain risks were collected and analyzed by term frequency, correlation, and bi-gram analysis. Topic modelling was also used to assist the analysis. Implementation of these text mining methods enabled the generation of a global supply chain risk categorization with a total of seven risk types and a dictionary of risk factor

terms. The risk types contain political, environmental, financial, supply and demand, logistic, system, and operational aspects. This categorization and the associated underlying risk factors can improve decision maker knowledge of global supply chain risks. Third, after categorizing risk types and defining potential risk factors, sentiment analysis was implemented on online news articles to recognize the variation of supply chain risks. Google News API was utilized to collect articles related to Yageo® Corp., a Taiwanese world-leading passive component provider. The underlying risk factor dictionary was used for article collection, where the financial and operational aspect of risks were targeted. The overall corpus level sentiments, polarity, and percent positive were calculated; these metrics indicated that the firm had a relatively positive pattern throughout 2018. Moreover, the sentiment polarity score and the stock price of Yageo® Corp. are moderately (positive) correlated. This result implies that the overall sentiment of the articles related to financial and operational supply chain risk is able to capture the risk variation pattern, after comparing the sentiments with stock prices.

This research is one of the first to utilize texts from extant literature to develop a risk categorization hierarchy and conducting sentiment analysis on news articles to reveal risk variation pattern. The academic contribution of this study is that it enhances the related extant works through developing a global supply chain risk categorization and risk pattern recognition framework. The positive impact on society and industry is that thanks to the power of text mining, decision makers of enterprises may update global supply chain risk types automatically, enriching their understanding of potential risks. They can also be aware of the dynamically changing relative importance of risks from variation patterns, obtained by analyzing real-time online news articles. Accordingly, further risk mitigation actions can be made. Overall, the proposed framework can improve enterprises' and researchers' understanding about global supply chain risks and further provide guidance in risk management conundrums.

The future extensions of this research will focus on analyzing the dependencies among risk types. Understanding the relationship among different risk types would significantly improve the global supply chain risk categorization and management. A large number of companies should be analyzed in the risk variation pattern recognition phase, preferably including co-dependent stakeholders in a supply chain. Their connections, responses and impacts to different risk events would also affect the risk management process.

Appendix

Article No.	Title	Authors (year)	Keywords	Journal
1	Supply chain management: a review of implementation risks in the construction industry	Aloini et al. (2012)	Supply chain management; Construction industry; Risk management	Business Process Management Journal
2	Approaches to managing global sourcing risk	Christopher et al. (2011)	Globalization; Risk management; Risk analysis	Supply Chain Management: An International Journal
3	Supply chain risk management: present and future scope	Ghadge et al. (2012)	Supply chain management; Risk management; Supply chain risk management; Systematic literature review; Text mining	The International Journal of Logistics Management
4	Supply chain risk management: a literature review	Ho et al. (2015)	Supply chain risk management; Risk types; Risk factors; Risk management methods; Literature review	International Journal of Production Research
5	Integrated fuzzy multi criteria decision making method and multi-objective programming approach for supplier selection and order allocation in a green supply chain	Kannan et al. (2013)	Green supply chain management (GSCM); Supplier selection; Multi-objective linear programming (MOLP); Maxi-min method; Order allocation	Journal of Cleaner Production
6	Global supply chain risk management strategies	Manuj and Mentzer (2008)	Risk management; Supply chain management; International business	International Journal of Physical Distribution & Logistics Management
7	Defining supply chain management	Mentzer et al. (2001)	N/A	Journal of Business Logistics
8	Product architecture and supply chain design: a systematic review and research agenda	Pashaei and Olhager (2015)	Product design; Systematic literature review; Platform; Integral, Modular, Supply chain design	Supply Chain Management: An International Journal

9	A decision framework for assessment of risk associated with global supply chain	Soni and Kodali (2013)	Risk management; Supply chain management; Decision making, Assessment; PROMETHEE; Goal programming; Global; Facility location	Journal of Modelling in Management
10	Identifying risk issues and research advancements in supply chain risk management	Tang and Musa (2011)	Supply Chain; Risk Management; Citation/Co-citation Analysis	International Journal of Production Economics
11	Strategic production-distribution models: A critical review with emphasis on global supply chain models	Vidal and Goetschalckx (1997)	Distribution; Strategic production-distribution models; Global supply chain modeling; Global logistics systems	European Journal of Operational Research

References

[1] Kannan, D., Khodaverdi, R., Olfat, L., Jafarian, A., & Diabat, A. (2013). Integrated fuzzy multi criteria decision making method and multi-objective programming approach for supplier selection and order allocation in a green supply chain. *Journal of Cleaner production, 47*, 355-367.

[2] Mentzer, J. T., DeWitt, W., Keebler, J. S., Min, S., Nix, N. W., Smith, C. D., & Zacharia, Z. G. (2001). Defining supply chain management. *Journal of Business logistics, 22*(2), 1-25.

[3] Barry, J. (2004). Supply chain risk in an uncertain global supply chain environment. *International Journal of Physical Distribution & Logistics Management, 34*(9), 695-697.

[4] Park, K., Kremer, G. E. O., & Ma, J. (2018). A regional information-based multi-attribute and multi-objective decision-making approach for sustainable supplier selection and order allocation. *Journal of Cleaner Production, 187*, 590-604.

[5] Koufteros, X., Vonderembse, M., & Jayaram, J. (2005). Internal and external integration for product development: the contingency effects of uncertainty, equivocality, and platform strategy. *Decision Sciences, 36*(1), 97-133.

[6] Christopher, M., Mena, C., Khan, O., & Yurt, O. (2011). Approaches to managing global sourcing risk. *Supply Chain Management: An International Journal, 16*(2), 67-81.

[7] Heckmann, I., Comes, T., & Nickel, S. (2015). A critical review on supply chain risk–Definition, measure and modeling. *Omega*, *52*, 119-132.

[8] Braithwaite, A. (2003). The supply chain risks of global sourcing. *LCP consulting*.

[9] Aloini, D., Dulmin, R., Mininno, V., & Ponticelli, S. (2012). Supply chain management: a review of implementation risks in the construction industry. *Business Process Management Journal, 18*(5), 735-761.

[10] Ghadge, A., Dani, S., & Kalawsky, R. (2012). Supply chain risk management: present and future scope. *The International Journal of Logistics Management, 23*(3), 313-339.

[11] Ponis, S. T., & Ntalla, A. C. (2016). Supply chain risk management frameworks and models: a review. *International Journal of Supply Chain Management*, *5*(4), 1-11.

[12] Yan, W., He, J., & Trappey, A. J. (2019). Risk-aware supply chain intelligence: AI-enabled supply chain and logistics management considering risk mitigation.

[13] Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: state of the art and future research directions. *International Journal of Production Research*, *57*(7), 2179-2202.

[14] Mitchelstein, E., & Boczkowski, P. J. (2009). Between tradition and change: A review of recent research on online news production. *Journalism, 10*(5), 562-586.

[15] Gupta, V., & Lehal, G.S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60-76.

[16] Meixell, M. J., & Gargeya, V.B. (2005). Global supply chain design: A literature review and critique. T*ransportation Research Part E: Logistics and Transportation Review, 41*(6), 531-550.

[17] Tang, O., & Musa, S.N. (2011). Identifying risk issues and research advancements in supply chain risk management. *International Journal of Production Economics, 133*(1), 25-34.

[18] Ho, W., Zheng, T., Yildiz, H., & Talluri, S. (2015). Supply chain risk management: a literature review. *International Journal of Production Research, 53*(16), 5031-5069.

[19] Chopra, S., Sodhi, M.S. (2004). Managing risk to avoid supply-chain breakdown. *MIT Sloan Management Review, 46*, 53–62.

[20] Cucchiella, F., Gastaldi, M. (2006). Risk management in supply chain: A real option

approach. *Journal of Manufacturing Technology Management, 17*, 700–720.

[21] Tummala, R., Schoenherr, T. (2011). Assessing and managing risks using the supply chain risk management process (SCRMP). *Supply Chain Management: An International Journal, 16*, 474–483.

[22] Tsai, M. C., Liao, C. H., & Han, C. S. (2008). Risk perception on logistics outsourcing of retail chains: model development and empirical verification in Taiwan. *Supply Chain Management: An International Journal, 13*(6), 415-424.

[23] Trkman, P., & McCormack, K. (2009). Supply chain risk in turbulent environments—A conceptual model for managing supply chain network risk. *International Journal of Production Economics, 119*(2), 247-258.

[24] Kayis, B., & Dana Karningsih, P. (2012). SCRIS: A knowledge-based system tool for assisting manufacturing organizations in identifying supply chain risks. *Journal of Manufacturing Technology Management, 23*(7), 834-852.

[25] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68.

[26] Power, D.J. (2016). Data science: supporting decision-making. *Journal of Decision systems*, *25*(4), 345-356.

[27] Van der Aalst, W.M. (2014). Data scientist: The engineer of the future. In *Enterprise interoperability VI* (pp. 13-26). Springer, Cham.

[28] Waller, M.A., & Fawcett, S.E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77-84.

[29] Sanders, N.R. (2016). How to use big data to drive your supply chain. *California Management Review*, *58*(3), 26-48.

[30] Nguyen, T., Li, Z. H.O.U., Spiegler, V., Ieromonachou, P., & Lin, Y. (2018). Big data analytics in supply chain management: A state-of-the-art literature review. *Computers & Operations Research*, *98*, 254-264.

[31] Choi, Y., Lee, H., & Irani, Z. (2018). Big data-driven fuzzy cognitive map for prioritising IT service procurement in the public sector. *Annals of Operations Research*, *270*(1-2), 75-104.

[32] Huang, Y. Y., & Handfield, R.B. (2015). Measuring the benefits of ERP on supply management maturity model: a "big data" method. *International Journal of Operations & Production Management*, *35*(1), 2-25.

[33] Wu, K. J., Liao, C. J., Tseng, M. L., Lim, M. K., Hu, J., & Tan, K. (2017). Toward sustainability: using big data to explore the decisive attributes of supply chain risks and uncertainties. *Journal of Cleaner Production*, *142*, 663-676.

[34] Chiu, M. C., & Lin, K. Z. (2018). Utilizing text mining and Kansei Engineering to support data-driven design automation at conceptual design stage. *Advanced Engineering Informatics*, *38*, 826-839.

[35] Formica, A. (2006). Ontology-based concept similarity in formal concept analysis. *Information sciences*, *176*(18), 2624-2641.

[36] Chang, D., & Chen, C. H. (2015). Product concept evaluation and selection using data mining and domain ontology in a crowdsourcing environment. *Advanced Engineering Informatics*, *29*(4), 759-774.

[37] Trappey, C. V., Trappey, A. J., Peng, H. Y., Lin, L. D., & Wang, T. M. (2014). A knowledge centric methodology for dental implant technology assessment using ontology based patent analysis and clinical meta-analysis. *Advanced Engineering Informatics*, *28*(2), 153-165.

[38] Govindarajan, U. H., Trappey, A. J., & Trappey, C. V. (2019). Intelligent collaborative patent mining using excessive topic generation. *Advanced Engineering Informatics*, *42*, 100955.

[39] Papadopoulos, T., Gunasekaran, A., Dubey, R., Altay, N., Childe, S. J., & Fosso-Wamba, S. (2017). The role of Big Data in explaining disaster resilience in supply chains for sustainability. *Journal of Cleaner Production*, *142*, 1108-1118.

[40] Kim, D., & Kim, S. (2017). Sustainable supply chain based on news articles and sustainability reports: Text mining with Leximancer and DICTION. *Sustainability*, *9*(6), 1008.

[41] Khan, M.N., Akhtar, P., & Merali, Y. (2018). Strategies and effective decision-making against terrorism affecting supply chain risk management and security: A novel combination of triangulated methods. *Industrial Management & Data Systems*, *118*(7), 1528-1546.

[42] Su, C.J., & Chen, Y.A. (2018). Risk assessment for global supplier selection using text

mining. *Computers & Electrical Engineering*, *68*, 140-155.

[43] Song, B., Yan, W., & Zhang, T. (2019). Cross-border e-commerce commodity risk assessment using text mining and fuzzy rule-based reasoning. *Advanced Engineering Informatics*, *40*, 69-80.

[44] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, *28*(2), 15-21.

[45] Chae, B. K. (2015). Insights from hashtag# supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, *165*, 247-259.

[46] Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458-464.

[47] Feinerer, I., & Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2.

[48] Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software*, *1*(3), 37.

[49] Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30.

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

[51] Murzintcev Nikita (2019). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.0. https://CRAN.R-project.org/package=ldatuning

[52] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

[53] Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. *In Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.

[54] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

[55] Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61-84.

[56] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

[57] Rangel, D. A., de Oliveira, T. K., & Leite, M. S. A. (2015). Supply chain risk classification: discussion and proposal. *International Journal of Production Research*, *53*(22), 6868-6887.

[58] Giannakis, M., & Papadopoulos, T. (2016). Supply chain sustainability: A risk management approach. *International Journal of Production Economics*, *171*, 455-470.

[59] Xu, M., Cui, Y., Hu, M., Xu, X., Zhang, Z., Liang, S., & Qu, S. (2019). Supply chain sustainability risk and assessment. *Journal of Cleaner Production*, *225*, 857-867.