Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 00 (2019) 000–000

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

25th International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing
August 9-14, 2019 | Chicago, Illinois (USA)

Applying Text-mining Techniques to Global Supply Chain Region Selection: Considering Regional Differences

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

The concept of the global supply chain is built upon the mindset of competitive advantages, and it results in several benefits to the manufacturing process, such as increasing flexibility and operational cost-down. Nevertheless, there are still plenty of risks differing among different regions or nations under a global supply chain, and this increases the uncertainty of the supply chain and the possibility of disruption. This paper tackles this issue through a text-mining based global supply chain regional risk identification framework that analyzes literature to identify potential regional global supply chain risks in a supplier region selection process. Text mining was implemented to collect and analyze the existing previous studies related to regional-related global supply chain risks. The results of text mining are represented in a document-term matrix that can further calculate the term-frequency (TF) and term frequency-inverse document frequency (TF-IDF) for the correlation analysis of terms related to supply chain risks. Those indicators provided information about whether the term-related content is vital in the studied literature, and further identified potential regional risk factors. A total of eight regional risk factors were extracted and organized. The proposed framework can objectively analyze geographical differences and provide enterprises with a global supply chain with intuitive guidance on region selection.

Keywords: Global supply chain; Supplier region selection; Risk analysis; Text mining

* Corresponding author. Tel.: +1-515-817-3537
E-mail address: cchu@iastate.edu

1. Introduction

The trend of the global supply chain has risen because it can achieve competitive advantages [1]. A global supply chain consists of a variety of stakeholders, from product design, manufacturing, assembly, to the end-customer. All stakeholders from all over the world participating in the supply chain network can be benefitted from it in terms of cost reduction, manufacturing flexibility, customer satisfaction, etc. [2]. However, there are still critical factors that can disrupt a global supply chain due to the complex environments it includes. Factors related to regional differences such as political changes, government policy, tax regulation, and natural disasters are several main examples of uncertainties that might harm the condition of a global supply chain [3] [4].

Another way to describe the global supply chain is as an integration of internal and external stakeholders, and this integration enhances the competencies of the stakeholders in product innovation, manufacturing agility, and other operational capabilities [5]. However, risks and difficulties still exist in building a robust global supply network due to the differences in multiple tiers of the system and different nations. Global sourcing risks invovle supply risk, environmental and sustainability risk, process and control risk, and demand risk [6]. Most of these risks are relatively more comfortable to be quantitatively assessed in the area of the domestic supply chain. However, environmental and sustainability risks are challenging to be evaluated because they are often affected by regional political, social, and economic uncertainties. Therefore, it is crucial to understand these geographical global supply chain risk factors to prevent the global supply chain from collapsing [6].

Paying attention to the above issues, the research question of this study are addressed as follows: What are significant global supply chain risks related to regional differences, and how do we identify them before disruptions occur? Most of the previous studies used case studies or literature reviews to categorize global supply chain risk; some also applied statistical or operational-modeling in problems of supply chain risk management [7] [8] [9]. A data-oriented method can help directly recognize regional-specific global supply chain risks.

Since there is plenty of information about geographical differences and region-related risks discussed in literature review articles and business magazines, a text-oriented, unstructured data analysis method, known as text mining, is required. Text-mining can analyze a large number of documents and extract valuable information such as term frequency, correlations, and topic clustering. Its exceptional capability of analyzing unstructured text data makes it useful in numerous fields of operations [10]. It is also extensively applied in knowledge discovery and management. According to the capability of text mining and the characteristics of geographical data, this study addresses the research questions through a risk identification process based on text mining to better understand uncertain global supply chain risks.

2. Literature review

Throughout recent decades, traditional supply chains have merged with globalization and new information management systems to integrate suppliers worldwide with their current business model [11]. Even so, there still are several issues that remain to be solved. Differences among regions such as taxes, exchange rates, and potential natural disasters add complexity to global supply networks. [12]. Besides, cultural differences and political environments increase the difficulty of maintaining a robust global supply chain [11]. Therefore, understanding regional risks is crucial in global supply chain management.

Previous studies discussed region related risks to some extent. Besides the risk classification proposed by Christopher et al. [11], global supply chain risks can also be decomposed into macro and micro risks [13]. Risk factors such as natural disasters, political threats, and local regulations are examples of macro risks, and risks related to the demand, supply, and manufacturing are in the category of micro risks [14] [15] [16].

Region related risk assessment methods are also important. Aloini et al. proposed that the first two steps of risk management are risk identification and risk quantification [8]. The risk identification focuses on determining the scenarios that would disrupt a global supply chain and exploring the degree of impact that disruption would bring to the system. The risk quantification emphasizes the process of assessing the severity, occurrence, and detection difficulties of a risk scenario. For example, Tsai et al. employed the analytic hierarchy process (AHP) to calibrate risk on outsourcing retail chains in Taiwan [17]. Trkman and McCormack took supplier characteristics, performance evaluation, operation environments into account in a risk identification conceptual model [18]. Kayis and Karningsih applied a knowledge-based tool to assist in supply chain risk identification [19].

Previous studies on risk identification in the global supply chain area demonstrate the importance of risk

management to a robust global supply chain. However, most of the identification model is in a qualitative or case-by-case fashion. The demand for a data-driven method able to develop a comprehensive global supply chain risk categorization by analyzing geographical difference is existing.

3. Methodology

In this paper, we have designed a method to identify regional risk factors to take into account while improving or maintaining global supply operations. Fig 1 summarizes the proposed method in this study. Text preprocessing and result analysis were completed with the assist of the R statistical programming language [20]. Package tm and tidytext are the primary toolkits used for data analysis and visualization [21] [22].

First, a corpus (a collection of 11 articles about global supply chain region related risks) was built in a folder and called to Rstudio for text data pre-processing. The 11 articles we selected in the corpus are from academic journals. They are well-cited, and the content is highly related to global supply chain risk management. A corpus in text-mining can have a variety of forms, such as pdf or Word files. There are several ways to transform the unstructured texts into structured data (i.e., txt file) in tm package. After the corpus setup, several steps are required to pre-process the documents. Generally, in an article, there are plenty of words and information. However, not all of them are needed to study a specific interest. For example, some of the terms are used to polish the idea the authors hope to present to their readers. Therefore, pre-processing and extracting essential words is vital while conducting text-mining. The content transformer function provided in tm package is capable of eliminating excess words by using some arguments such as removeNumbers, removePunctuation, etc. By using this function, we converted all texts into lower case and removed all numbers, punctuations, and whitespaces. Furthermore, each word in the corpus was transformed into a word stem using the SnowballC package [23]. According to the above data pre-processing, further text-mining analysis can proceed based on the clean and tidy corpus.

Second, after organizing and tuning the corpus, a document term matrix was generated. In this matrix, each row represents a document within the corpus, and each column represents a stemmed term that appears in the corpus. Term frequency (TF) can be calculated by getting the column sum. Frequency analysis was conducted by summing up the counts of each term and finding the most import terms. Correlations between words and phrases were also calculated to explore the relationships between words. After frequency and correlation analysis among critical terms, a set of global supply chain regional risk factors was categorized.

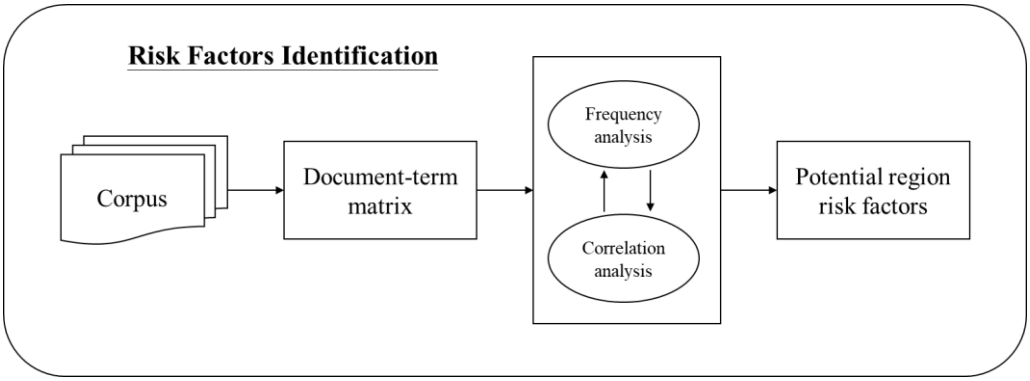

Fig. 1. Global supply chain regional risk identification.

4. Results

The corpus in this study contained 11 journal articles from several journals; they are all related to the field of global supply chain risk management. After data pre-processing (i.e., remove numbers, punctuations, stop words), the corpus consisted of 4290 stemmed words. We then converted the corpus to the document-term matrix (Table 1).

Table 1. Document-term matrix (partial).

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

Notice that there are quite a few 0s in the matrix. The function removeSparseTerms was used to simplify the matrix into essential 199 terms left. Frequency analysis was then conducted to obtain information such as TF and TF-IDF, and the results are shown in Table 2 and Fig 2. The terms set frequently appearing throughout the corpus is reasonable, because the articles selected in the corpus is mainly about global supply chain risk management. Although the output is obviously related to the selection process of the corpus, the visualization results provided guidance during the first phase of the analysis.

Table 2. Top eight frequent terms among all documents.

Stemmed term / original words	Frequency
"risk" / risks	1365
"suppli" / supplied or supplies	1208
"chain"/ chain	1010
"manage" / manage, manager, managerial, and management	599
"product" / product, productivity, and production	526
"model" / model and modeling	467
"supplier" / supplier	465
"global"/ global and globalization	321

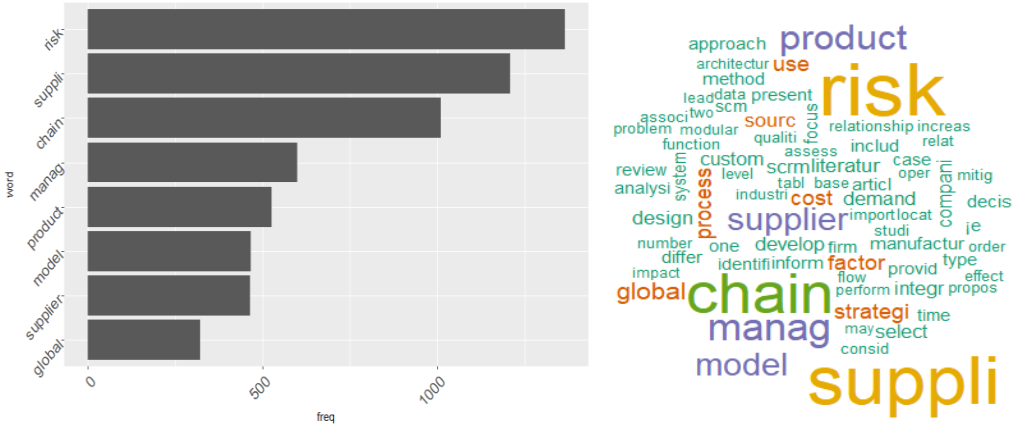

Fig 2. (a) Frequent terms (TF > 250). (b) Word cloud.

In order to discover potential risk factors by analyzing the corpus, correlations between terms (especially with term "risk") were analyzed. Table 3 demonstrates some of the words are having relatively high associations with "risk" (with a correlation threshold of 0.6). The result shows that the term "risk" is highly correlated with the word "terror" and "war", and this can indicate that terrorism or terrorist and war are significant risk factors. These could be classified as regional risk factors because terrorist events or war generally happen in a particular area during a specific time period. Terms possibly related to environment or climate change such as "natur", "disast", "earthquak", and "tsunami" are also highly correlated to risk. When a natural disaster such as earthquakes and tsunamis happen in a particular

region, the global supply chain would be susceptible to risk due to the damage to the infrastructure and economics. Other than these, terms “competitor”, “retail”, “bullwhip”, and “demand” could also be risk-related terms, and they are usually used to describe the system-wide risk of global supply chains. This result is validated that most of the global supply chain risk management studies considered in this research regarded terms having a correlation higher than 0.75 as common risk factors such as terrorism and natural disaster.

Table 3. Terms having high correlations with “risk”.

Stemmed term / original words	Correlation
“terror” / terrorism and terrorist	0.88
“failur” / failure	0.84
“occurr” / occurrence	0.83
“maritim” / maritime	0.81
“disast” / disaster	0.80
“bullwhip” / bullwhip	0.79
“competitor” / competitor	0.78
“demand” / demand	0.78
“mitig” / mitigate and mitigation	0.78
“war” / war	0.76
“retail” / retail and retailer	0.76
“insur” / insurance	0.75
“natur” / nature and natural	0.74
“catastroph” / catastrophy	0.73
“uncertain” / uncertain and uncertainty	0.71
“substitut” / substitute	0.69
“bankruptci” / bankruptcy	0.66
“earthquak” / earthquake	0.65
“transpar” / transparent and transparency	0.64
“tsunami” / tsunami	0.60

Knowing how often a term comes before and after the word “risk” is also helpful in identifying the potential risk factors. The corpus was converted to tidy-verse text form for n-gram analysis (n represents the number of words consecutively appear together). Usually, in text mining, a document-term matrix is built on the basis of the one-column-one-term matrix. By using tidytext package, an n-gram analysis can be implemented. A bi-gram analysis was performed to calculate the frequencies of terms that come before and after the word “risk”. Different from the analysis in Table 3, the analysis here strictly focuses on consecutive terms as seen in Fig 3. The phrases “risk factor” and “risk management” appear more than 105 times and 90 times in Fig 3, respectively. Other terms that came after risk also provided information on the subject discussed in the articles while talking about “risk”. Moreover, terms that came directly before risk may give us more intuitive knowledge of the potential risk factors. In Fig 3 (b), there are several frequent terms that came before “risk” that demonstrate some connections with risk factors, such as “sourc” (source), “suppli” (supply), “chain”, “financi” (financial), “demand”, “manufactur” (manufacture), etc. Fig 4 visualizes a bi-gram network that can easily show relationships between the bi-grams that occur more than 20 times.

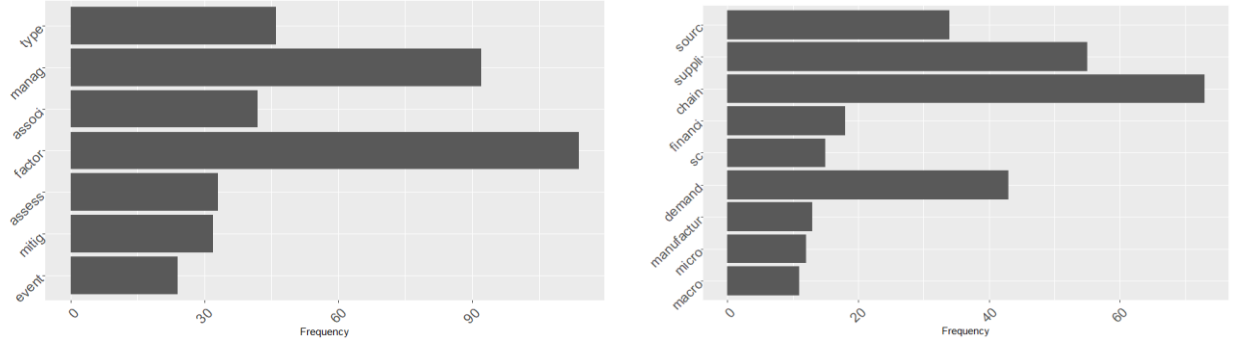

Fig 3. (a) Frequencies of the term come after the term “risk” (Freq. > 20). (b) Frequencies of the term come before the term “risk” (Freq. > 10).

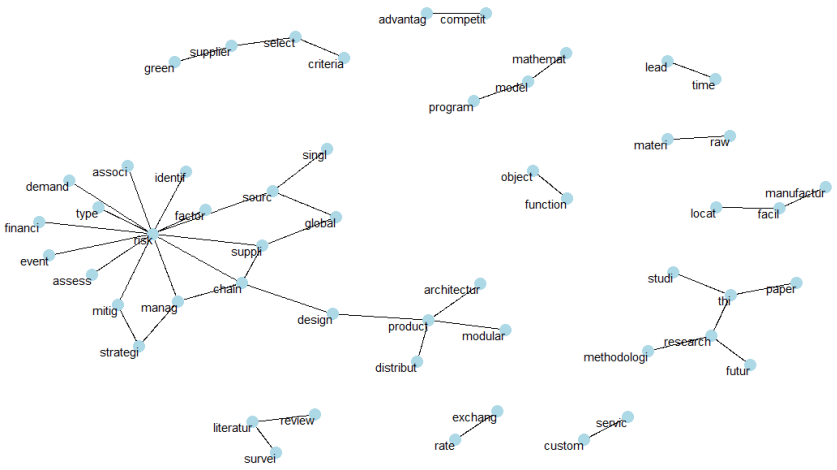

Fig 4. The bi-gram network

By taking frequency and correlation analysis into consideration, eight potential global supply chain risk factors related to regional differences were discovered (Table 4). These risk factors are political risk, logistic risk, environmental risk, financial risk, supply risk, demand risk, system risk, and information risk. The names of the risk types could vary vastly depending on each person's understanding; however, the critical point here is what attributes define them. In previous studies, researchers studied many articles and case studies to explore the relationship between global supply chains and the underlying side-effects to manage while organizing the risk factors. Nevertheless, by implementing the proposed method that uses text-mining, regional risk factors could be relatively easily categorized and discovered by analyzing related articles. Although this output of geographical risk factor categorization might not be comprehensive enough to cover all kinds of risks that relate to a global supply chain, this study demonstrates the potential of the proposed method and provides valuable insights for both academia and industries for future global supply chain risk research.

Table 4. Potential global supply chain region risk factors.

Stemmed term / original words	**Regional risk factor**
"terror" / terrorism and terrorist "war" / war	*Political risk*
"maritim" / maritime "distribut" / distribution	*Logistic risk*
"disast" / disaster "natur" / nature and natural "catastroph" / catastrophy "tsunami" / tsunami "earthquak" / earthquake	*Environmental risk*
"bankruptci" / bankruptcy "financi" / financial	*Financial risk*
"suppli" / supply and supplier "substitut" / substitute and substitution "sourc" / source "manufactur" / manufacturer, manufacture, and manufacturing	*Supply risk*
"demand" / demand	*Demand risk*
"retail" / retail and retailer	*System risk*

"uncertain" / uncertain and uncertainty	
"sc" / supply chain	
"chain" / chain	
"transpar" / transparent and transparency	*Information risk*

5. Conclusion

The global supply chain is an essential concept for companies. The resilience of the supply chain is also a crucial area for global supply chain management. In order to enhance the robustness and performance of supply chains, risk analysis should be implemented. However, in the context of globalization, there are numerous regional differences that would influence the global supply chain resilience. Therefore, regional risk factor identification is essential and should be taken into consideration in supplier selection process.

This study incorporates text mining to identify potential global supply chain risk factors from related extant studies. First, 11 major articles were included in the corpus as an input of the text mining approach. Second, after pre-processing the corpus, the document-term matrix was created. This step converted unstructured text data into structured data for further information extraction. Third, frequency analysis and correlation analysis were conducted to discover the relationships among terms. The visualization tool in R was used throughout the research, which provided with straightforward comparison and valuable insights. Finally, eight potential regional risk factors were extracted and organized based on the former analysis. This set of geographical risk factors could provide guidance for companies in region selection while building or redesigning their global supply chain.

There are limitations to this research. There were only 11 research articles in the corpus, which might give us limited information. If the size of the corpus increases, the results may be more comprehensive. Correlated topic models could also be implemented to cluster the corpus. This descriptive analysis might further improve risk factor identification. Despite these limitations, this study can be as a basis to improve a region selection process in global supply chain design.

As for future direction, risk evaluation of a region can be implemented by applying sentiment analysis. Sentiment analysis is based on counting the scores given to each term according to their tendencies toward positivity or negativity. When there is a new article discussing global supply chain risk or any regional differences or events of a particular region, the proposed set of risk factors can help determine which factors were potentially addressed in that article. The enterprises can then determine the severity of the risk in that region based on the overall sentiment scores.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.
[13] 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.
[14] Chopra, S., Sodhi, M.S. (2004). Managing risk to avoid supply-chain breakdown. *MIT Sloan Management Review, 46*, 53–62.
[15] Cucchiella, F., Gastaldi, M. (2006). Risk management in supply chain: A real option approach. *Journal of Manufacturing Technology Management, 17*, 700–720.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[21] Feinerer, I., & Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2.
[22] 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.
[23] Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. *R package version 0.5*, *1*.