SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

SEMARAK ILMU
Journal of
ADVANCED RESEARCH IN APPLIED SCIENCES AND ENGINEERING TECHNOLOGY
ISSN 2462 - 1943

Role of Text Mining in Extracting Valuable Information from Text Data

Zatul Alwani Shaffiei[1,*], Amir Syafiq Syamin Syah Amir Hamzah[2,3], Shaikh Mariyam Harunor Rashid[1], Naoki Oshima[4]

[1] Department of Electronic Systems Engineering (ESE), Malaysia–Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[2] Department of Management of Technology (MoT), Malaysia–Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[3] Intellectual Property and Innovation Management (IPIM) iKohza, Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[4] Graduate School of Management of Innovation and Technology, Yamaguchi University, Japan

ARTICLE INFO

Article history:
Received 2 February 2023
Received in revised form 5 August 2023
Accepted 15 August 2023
Available online 3 September 2023

Keywords:
Text mining; Word cloud; Dendrogram; Co-occurrence network; Euclidean distance

ABSTRACT

Text mining has become a popular field with the rapid development of information technology and the extensive amounts of unstructured text data such as web pages, social network sites and technical documentations. This data contains a lot of information, which is extremely difficult to deal with the huge number and various forms. Extracting and analysing important information from massive data, for example in automotive industries has become our major problem. The main aim of text mining is to extract important information from massive text data that are difficult to handle manually with error-free. In this paper, the fundamental concept is based on Euclidean distance in finding the similarity between words. Finally, a set of data is used to describe the similarities, distances and frequencies between several words. Word cloud, bar plot, dendrogram and co-occurrence network are also presented to illustrate the behaviour of the text data.

1. Introduction

Individuals and organizations generate enormous amounts of data every day. Statistically 80%of existing text data is unstructured, therefore it's not searchable and therefore is not useful. Only 18% of organizations have taken advantage of this data [1]. The Internet is a melting pot of massiveamounts of text data and analysing all the information gathered if done manually is a lengthy and

complicated process. Organizing, categorizing, and capturing relevant information from raw unstructured data is a major concern and challenge for companies and text mining is crucial for this.

Text mining is a subdivision of data mining that converts unstructured text into structured data by using natural language processing (NLP), which allows our machine to understand human language, and statistical and machine learning techniques to process it automatically. Meaningful patterns are then identified from the data to extract useful information [2]. As an example, emails, social media posts, chats and surveys are unstructured text data and if a few people were tasked to sort through this

[*] *Corresponding author.*
E-mail address: zatulalwani.kl@utm.my

https://doi.org/10.37934/araset.32.1.263271

information and provide conclusions it might lead to errors and failure for the business not only because its time consuming and expensive, but also because it might be inaccurate and is impossible to scale. Text mining provides a reliable and cost- effective solution to this while simultaneously achieving scalability and accuracy in a smaller amount of time.

This paper is divided into several sections. The introduction explains the issue of text data, internet as well as some advantages of using NLP which has become our main focus. In the literature review, the text mining processes related to various industries have been discussed. It includes other studies that have obtained relevant methods to solve text mining problems in the past until the recent finding such as that conducted by Kushwaha *et al.,* [3]. Meanwhile, methodology discussed in depth the proposed solution in which R software is employed. It covers three main steps which are text pre-processing, text mining operations and post processing and provides useful functions in the form of packages which are stored in the *R* libraries. Finally, results and conclusion summarise the study with a conclusion, re-stating the contributions as well as some suggestions for future research.

2. Literature Review

Text mining is an amalgamation of three research areas: information retrieval, data mining and NLP (natural language processing). Information retrieval started in the 1960s and mainly dealtwith text retrieval. Principally, if a few keywords are provided, we can find related documents from a text collection. Web search engines, like Google, Bing and Yahoo are essentially information retrieval systems [4]. NLP started in the 1950s to make computers understand human language. Data mining uses structured data, but in the late 1990s, researchers began to use text as data, which gave rise to text mining. In the beginning text mining applied data mining and machine learning algorithms to analyse data. NLP techniques weren't used then, but in the last 10 years as the scope of text mining research expanded, natural language processing techniques like parsing, part-of-speech tagging, summarization, etc. were integrated [5].

Approximately 80% of data in the world is unstructured data and 63-70% of this data goes unused. Text mining is an efficient way to derive important conclusions from this data. This helps usimprove the decision making of organizations leading to more successful endeavours [6]. Text mining has impacted many industries such as automotive, healthcare and telecommunication, allowing them to cater products according to a customer's demands and make faster and better business decisions [7-9]. Companies have surveys, onlinereviews and social medial profiles and text mining is done on the data gathered here to analyse which product is more popular to increase its manufacture and use the comments as a benchmark to create better products in the future. It may also be used in risk management to provide insights on industry trends and financial markets by monitoring changes in sentiment by analysing reports and whitepapers. It is used to filter spam as text mining allows us to exclude emails from certain senders in our inboxes.

The National Centre for Text Mining (NaCTeM), operated by the University of Manchester in close collaboration with Tsuiji Lab, University of Tokyo, is the first publicly funded text mining centre in the world. It provides customised tools, research facilities and offers advice to academic community. The initial focus is text mining in the biological and biomedical sciences, but research has expanded into areas of social sciences [10]. While text mining is helpful, it comes with its own fair share of challenges. While it performs some tasks reasonably well, many other tasks need improvement in accuracy. To master text mining,one must have an in depth understanding of natural language. Although much research has been conducted in this field, the progress is not great. Currently text analysis techniques are mainly basedon statistical machine learning and data mining algorithms.

These methods don't provide true and accurate understanding which reduces accuracy of tasks. Another major issue is sentence structure may change the way the data is interpreted by the system

[11]. Conceptually, it seems promising and if done well and accurately could be beneficial to the industry, but that can be very challenging.

3. Methodology

The text analysis of each dataset in this study is done using *R* software. *R* is a free software that providesan extensive variety of statistical and graphical techniques. It is easy to learn and use and it produces well defined graphical plots for datasets. It can also perform several other functions such as statisticalanalysis, probability distribution and random number generation among others.

R provides useful functions in the form of packages which are stored in libraries. Each library contains tools, functions, and methods to achieve a certain goal. For instance, the library 'dplyr'is used for data manipulation, 'ggplot2' is used for data visualization in the form of pie charts, histograms, etc. The datasets used in this study are patent datasets from Derwent Innovations from Clarivate,a global leader in providing analytics to accelerate the process of innovation. It aims to improve the way the world creates, protects, and advances innovation.

Figure 1 shows the text mining process which can be divided into three main steps: text pre-processing, text mining operations and post processing [12]. Pre-processing involves data selection, classification, and normalization (transforming data into a more compatible form that allows different text mining techniques to be implemented). Next techniques like clustering, association rule detection, visualization, and terms frequency are applied on the normalized data. In the third step, alterations are made on the data through text mining functions like evaluation and choice of knowledge.

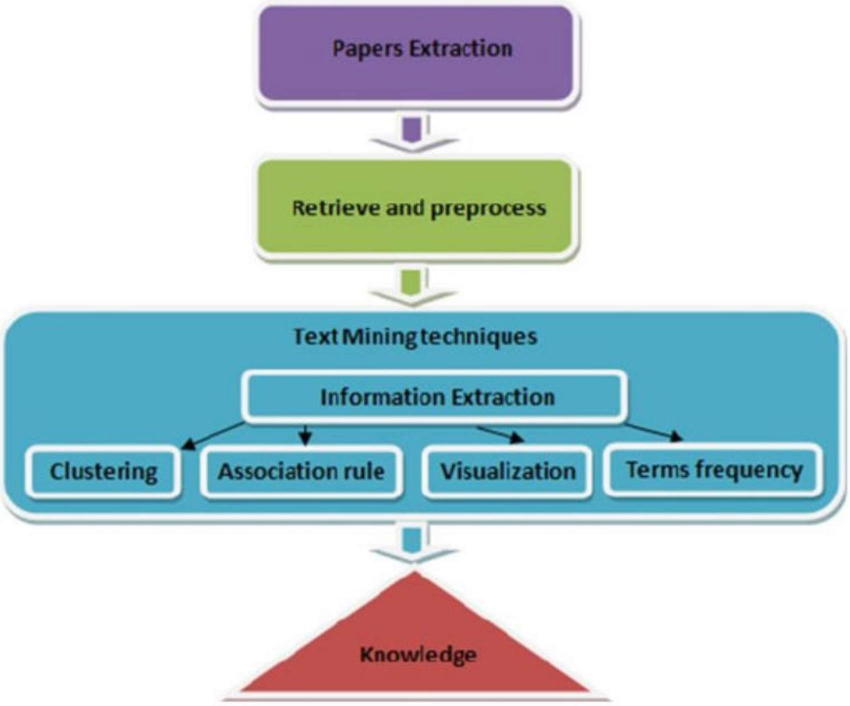

Fig. 1. Text mining processes [12]

4. Results and Discussion

In *R* software, before using the library, few packages should be installed in our system using code:

```
install.packages("ggplot2")
```

For each subsequent use, the library must be loaded in the beginning of the session using:

```
library("ggplot2")
```

Some of the libraries used for the demo in this paper are tm, tmap, wordcloud, dplyr, amongothers [13]. The dataset is copied into a text file and loaded into our session. The data is then convertedto a corpus to make data manipulation in the subsequent steps easier.

```
mydata <- Corpus(VectorSource(newdata))
```

Pre-processing is performed to filter the data and remove any numbers, symbols such as /, @, and common English stop words such as there, and the. Any specific word throughout the entiretext that has been considered as unimportant also can be removed. All the text is converted to lowercase to reduce redundancy and extra whitespace is removed. The code for performing pre- processing is:

```
mydata <- tm_map(mydata, removeWords, stopwords("english"))
```

A term document matrix is then created from the processed corpus. Then the text is analysedfor frequency and sorted in descending order and the top 20 records are used for data visualization.For the demonstration, a sample dataset of automotive industry for company A from Derwent Innovation has been used. By applying text mining, the strength of technology and expertise for particular company can be determined. Based on this sample dataset, there are four results that have been produced: bar plot, word cloud, dendogram and co-occurrence network.

4.1 Bar Plot

A simple data-visualization – a bar plot has been produced as shown in Figure 2. The X-axis showsthe words with most frequencies in descending order and the Y-axis shows the frequencies of the words. From this figure, “pipe” and “hose” word are the most frequently occurring terms which are 41 and 38 times respectively, while resin and metal occur 12 times. The following code is used to generate this result:

```
barplot(dtm_d [1:20,]$freq, las=2, names.arg = dtm_d[1:20,]$word, col
="lightblue", main ="Top 20 most frequent words", ylim=c(0,50),ylab =
"Word frequencies")
```

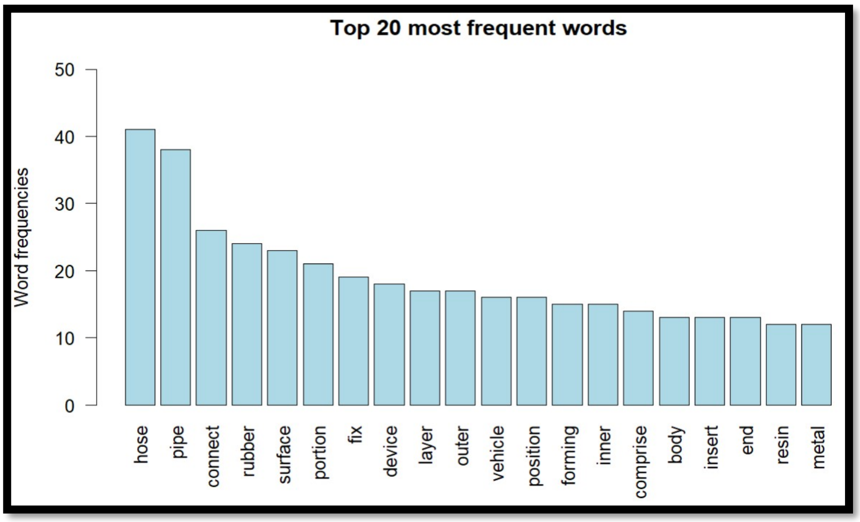

Fig. 2. Bar plot of 20 most frequently occurring words

4.2 Word Cloud

Next, by using the same sample dataset, word cloud also has been produced. A word cloud isa cluster of words that may be depicted in different sizes and colours to demonstrate a particular result in a dataset. In our demo, word clouds have been used to demonstrate frequencies of different words. The font size of the word is based on its frequency; the more frequently it appears, the larger the font size, as depicted in Figure 3. “pipe” and “hose” word have the highest frequency, thus the fontsize for these two words will be the largest.

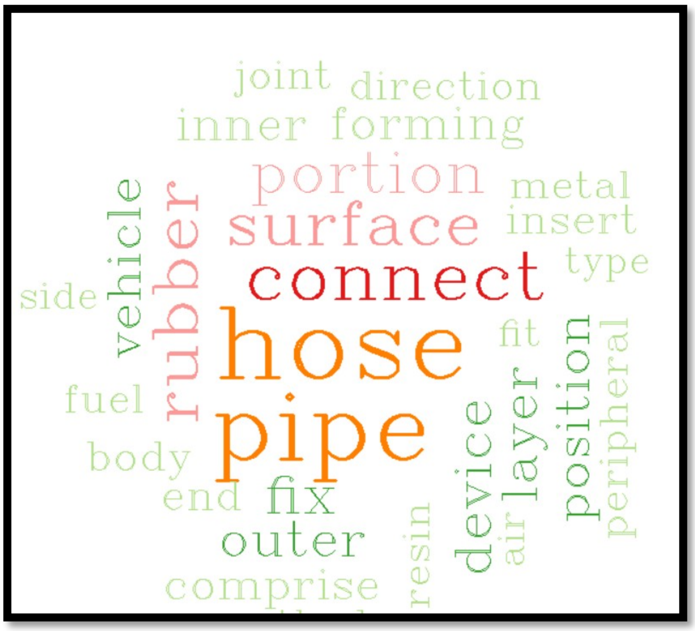

Fig. 3. Word cloud based on frequency of words

The following code is used to generate the result of word cloud in Figure 3:

```
    wordcloud (words=dtm_d$word, freq=dtm_d$freq, min.freq=12,
          max.words=100, random.order=FALSE, rot.per=0.40,
  colors=brewer.pal(9,"Paired"), vfont=c("sans serif","plain"))
```

It can also be customized to display words that can be differentiated based on depth of colour.If the shade is darker, then its more frequent, if it is less frequent, then the shade is lighter, as in Figure 4. The words pipe and hose are in the darkest shade of green while words like direction, resin and joint are in a lighter shade of green due to comparatively less frequency. The following code is used to generate the result of word cloud in Figure 4:

```
      wordcloud(words=dtm_d$word, freq=dtm_d$freq,
     min.freq=12,max.words=100, random.order=FALSE,
                      rot.per=0.40,
  colors=brewer.pal(8,"BuGn"), vfont=c("serif", "bold"))
```

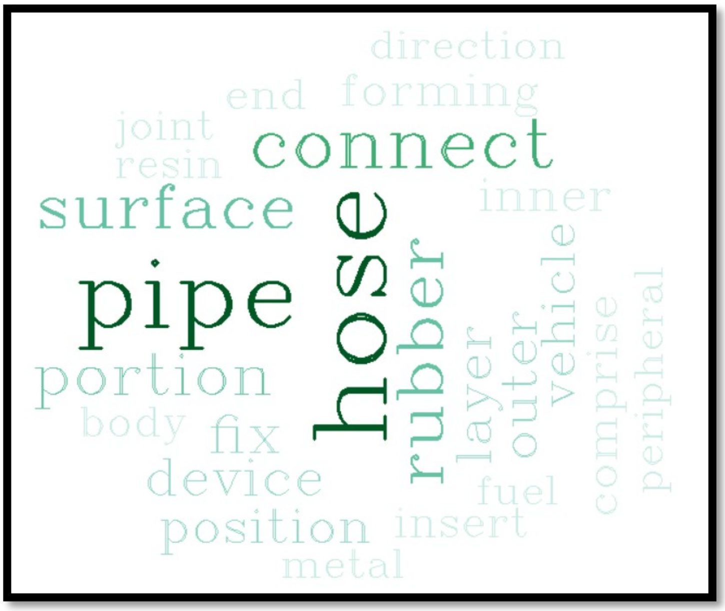

Fig. 4. Word cloud differentiating based on depth of color

4.3 Dendogram

Other than bar plot and word cloud, a dendrogram also can be generated based on the givendataset. Dendrogram is a tree diagram that shows hierarchical clustering (clustering that groups similar objects together).

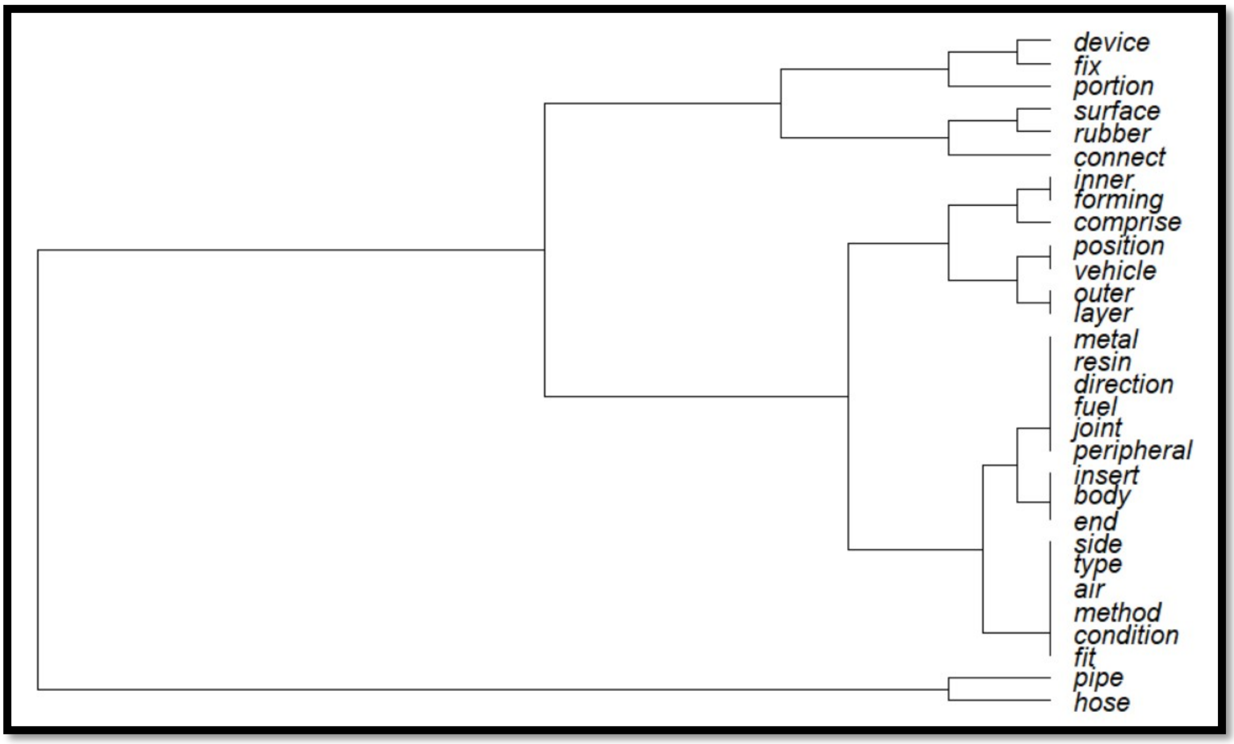

Fig. 5. Dendrogram from cluster analysis

In *R*, words can be clustered using the distance matrix [14]. First, a dissimilarity matrix is constructed; currently R provides three methods to perform this function which are: euclidean, manhattan and gower [15]. For this study, the ground calculation that has been used is Euclidean distance method [15]. In Figure 5, the top 30 most frequent words have been presented. The function ‘hclust’provides several methods for cluster analysis such as complete, median, single,centroid among others. The following code produces the result in Figure 5:

```
d <- dist(head(dtm_d,30), method = "euclidean") hc1 <- hclust (d,
method = "complete”) plot(as.phylo(hc1), cex = 0.8, label.offset =
0.5)
```

Words that are clustered together have comparatively similar frequencies. Therefore, wordslike device and fix have a closer frequency that is less on the scale and words like pipe and hose havea closer frequency that is higher on the scale. From this figure, the words metal, resin, direction, fuel,joint and peripheral also have the same frequency.

4.4 Co-occurrence network

The words that occur most frequently together using a co-occurrence network also can be determined [16-18]. Co-occurrence analysis identifies pairs of words that occur together in the text file. The frequency of the pair occurring together increases density and width of the line connecting the two words. It starts with creating a bigram and counting frequencies of all the pairs occurring together, then it is arranged in descending order, the pair with the highest frequency on top to ensure that data visualization is error free. The bigram is then plotted with the following code:

```
bigrams_counts %>% filter(n >= 5) %>% graph_from_data_frame() %>%
ggraph(layout = "fr") + geom_edge_link(aes(edge_alpha = n, edge_width
= n), color = "violetred2") +
geom_node_point(color = "midnightblue", size = 5) +
geom_node_text(aes(label = name), vjust = 1.8) +
ggtitle(expression(paste("Cooccurrence Network - ",
italic("Company A")))) + theme_void()
```

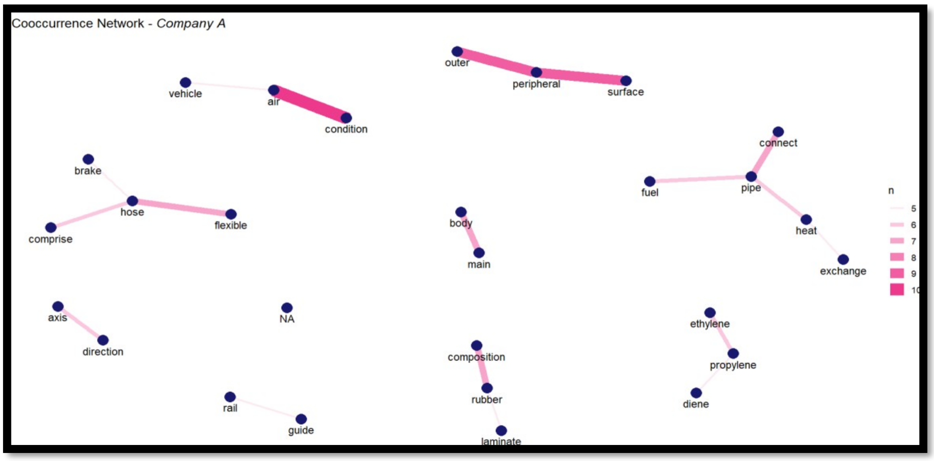

Fig. 6. Co-occurrence networks

As depicted in Figure 6, words like outer, peripheral and surface occur together more frequently, which can be considered as a main theme that has been focused. While for the word air, conditionand air occur together more frequent compared to vehicle and air which is occur together less frequent. Similar patterns can be detected in the above figure. *n* indicates the number of occurrences.

4. Conclusions

This study explored the extent of research done in the field of text mining which is covered on information retrieval, classification, and organization from unstructured data. Text mining allowsus to mine data for people's thoughts and opinions, genuine insights, and both positive and negative feedback. Text mining implementation has been unpopular in the past due to less research in the field. However, in recent years, text analytics adoption rates are increasing, and high technological advancements like automated text analysis can distinguish valuable data from unstructured information.

For future work, text mining can be used as a part of research in open innovation. The objective is to research patents owned by different organizations, to understand how they can participate in open innovation. It will help in understanding the impact of intellectual property management that considers patent risks on stock prices. However, in this study, there is a limitationthat should be improved from the results, in terms of removing unnecessary words that are not related with the focus theme.

References

[1] Smith, Tim, Ben Stiller, Jim Guszcza, and Tom Davenport. "Analytics and AI-driven enterprises thrive in the Age of With." *Deloitte Insights* (2019).

[2] Rybchak, Zoryana, and Oleh Basystiuk. "Analysis of methods and means of text mining ECONTECHMOD." *An International Quarterly Journal* 6, no. 2 (2017): 73-78.

[3] Kushwaha, Amit Kumar, Arpan Kumar Kar, and Yogesh K. Dwivedi. "Applications of big data in emerging management disciplines: A literature review using text mining." *International Journal of Information Management Data Insights* 1, no. 2 (2021): 100017. https://doi.org/10.1016/j.jjimei.2021.100017

[4] Giordano, Vito, Filippo Chiarello, Nicola Melluso, Gualtiero Fantoni, and Andrea Bonaccorsi. "Text and dynamic network analysis for measuring technological convergence: A case study on defense patent data." *IEEE Transactions on Engineering Management* (2021).

[5] Salloum, Said A., Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. "Using text mining techniques for extracting information from research articles." *Intelligent natural language processing: Trends and Applications* (2018): 373-397. https://doi.org/10.1007/978-3-319-67056-0_18

[6] Wang, Dakuo, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. "Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai." *Proceedings of the ACM on human-computer interaction* 3, no. CSCW (2019): 1-24. https://doi.org/10.1145/3359313

[7] Kim, En-Gir, and Se-Hak Chun. "Analyzing online car reviews using text mining." *Sustainability* 11, no. 6 (2019): 1611. https://doi.org/10.3390/su11061611

[8] Sun, Wencheng, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang. "Data processing and text mining technologies on electronic medical records: a review." *Journal of healthcare engineering* 2018 (2018). https://doi.org/10.1155/2018/4302425

[9] Yang, Kenneth CC, and Yowei Kang. "Framing national security concerns in mobile telecommunication infrastructure debates: A text mining study of Huawei." *Huawei Goes Global: Volume II: Regional, Geopolitical Perspectives and Crisis Management* (2020): 319-339. https://doi.org/10.1007/978-3-030-47579-6_14

[10] Rahaman, Tariq. "Discovering new trends & connections: current applications of biomedical text mining." *Medical Reference Services Quarterly* 40, no. 3 (2021): 329-336. https://doi.org/10.1080/02763869.2021.1945869

[11] Henriksson, Aron, Jing Zhao, Hercules Dalianis, and Henrik Boström. "Ensembles of randomized trees using diverse distributed representations of clinical events." *BMC medical informatics and decision making* 16, no. 2 (2016): 85-95. https://doi.org/10.1186/s12911-016-0309-0

[12] Yousuf, Hana, Asma Y. Zainal, Muhammad Alshurideh, and Said A. Salloum. "Artificial intelligence models in power system analysis." In *Artificial intelligence for sustainable development: Theory, practice and future applications*, pp. 231-242. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-51920-9_12

[13] Chen, Xieling, Haoran Xie, Fu Lee Wang, Ziqing Liu, Juan Xu, and Tianyong Hao. "A bibliometric analysis of natural

language processing in medical research." *BMC medical informatics and decision making* 18, no. 1 (2018): 1-14. https://doi.org/10.1186/s12911-018-0594-x

[14] Barter, Rebecca L., and Bin Yu. "Superheat: An R package for creating beautiful and extendable heatmaps for visualizing complex data." *Journal of Computational and Graphical Statistics* 27, no. 4 (2018): 910-922. https://doi.org/10.1080/10618600.2018.1473780

[15] Kumar, Ashish, and Avinash Paul. *Mastering text mining with R*. Packt Publishing Ltd, 2016.

[16] Silge, Julia, and David Robinson. *Text mining with R: A tidy approach*. " O'Reilly Media, Inc.", 2017.

[17] Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007. https://doi.org/10.1017/CBO9780511546914

[18] Barth, James R., Hemantha SB Herath, Tejaswini C. Herath, and Pei Xu. "Cryptocurrency valuation and ethics: a text analytic approach." *Journal of Management Analytics* 7, no. 3 (2020): 367-388. https://doi.org/10.1080/23270012.2020.1790046