**ZULFAQAR Journal of Defence Science, Engineering & Technology**
e-ISSN: 2773-5281
Vol. 5, Issue 2 (2022)
DOI: https://doi.org/10.58247/jdset-2022-0502-12
Journal homepage: https://zulfaqarjdset.upnm.edu.my

# A COMPARISON STUDY ON TEXT MINING AND SENTIMENT ANALYSIS FEATURES AND FUNCTIONS USING SAS ENTERPRISE MINER, PYTHON AND R

**Yi Fei Lee[a*], Angela Siew Hoong Lee[a], Zuraini Zainol[b]**

[a] School of Engineering and Technology, Sunway University, Bandar Sunway, 47500 Selangor, Malaysia
b Department of Computer Science, Faculty of Science and Defence Technology, National Defence University of Malaysia, Sg. Besi Camp, 57000 Kuala Lumpur, Malaysia

**ARTICLE INFO**

**ARTICLE HISTORY**
Received: 31-05-2022
Revised: 10-07-2022
Accepted: 30-08-2022
Published: 31-12-2022

**KEYWORDS**
Text mining
Text cleaning
Sentiment analysis
Phyton
SAS and R

**ABSTRACT**

T Twitter has allowed textual data to be collected using Text Mining and Sentiment Analysis techniques in the age of social media in which user-generated content becomes redundant. However, due to some inconsistencies, Text Cleaning plays an important role before Text Mining and Sentiment Analysis techniques can be conducted. Hence, this study is conducted to discover the capabilities of Text Cleaning, Text Mining and Sentiment Analysis in three different data mining tools: SAS® Text Miner (proprietary text mining tool), Python and R programming (open-source text mining tools). These data mining tools were used to conduct the Text Cleaning, Text Mining and Sentiment Analysis and their capabilities such as features, functions and characteristics were evaluated and investigated, to conduct this comparison study. All the proposed research objectives were met successfully even with the given limitation. A movie critique Dictionary is one of the major theoretical implications of this research. Based on our analysis and results, developers or educational practitioners can discover what is important and what is unimportant when conducting Text Mining and Sentiment Analysis. They will also obtain insights and guidance on how to conduct Text Mining and Sentiment Analysis using SAS Enterprise Miner, Python and R.

## 1.0 INTRODUCTION

Social media example Twitter, Instagram, Facebook and so on, have become a common repository of knowledge containing the state of mind of humans on various subjects since the transformation of Web 3.0 This is because social media platforms have enabled the social media users to broadcast their personal status, random thoughts and opinions regarding to any topic of their interest in a split second, as long as they are equipped with devices and connected with the Internet connection. This scenario has led to the growing volume of user generated data in social media. According to [1], user generated content contains valuable opinions (corpus) from the public. If the content is mined and analyzed properly, it can turn out to become useful knowledge and beneficial for variety applications that require text mining and sentiment analysis results.

Text Mining or text analytics is a technique used for discovering interesting patterns and trends from gigantic amount of text data [2]. Numerous recent studies with a wide variety of applications have applied text mining method such as Sentiment Analysis to analyze and discover insights from social media platforms [3] – Twitter, Instagram, Facebook, Reddit, Snapchat, etc. For instant, SA has been applied to mine public opinion on social media platforms about customers' feedback and their experiences on products [4] and services [5-7], predicting elections and candidates popularity [8-9], analyzing cyberbully behavior and patterns [10-12] and health public information [13-14], and many more. Organizations are constantly discussing the preference of one software tool over another considering factors like employee's current technical skills, learning ability and the graphical capability of the tool [15]. There are

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my

a lot of journal articles discussing on the use of Text Mining and Sentiment Analysis on different fields, comparing different Data and Text Mining Tools, but only few of them had discussed on the Sentiment Analysis features present in different tools and very few discussed on detail on the level of Sentiment Analysis the tools can perform [16-25]. Based on the above discussions, it is worthy to conduct an overview and a comparative analysis of SAS Enterprise Miner (proprietary text mining tool), Python and R programming (open-source text mining tools) respectively based on their capabilities to answer the following research questions:

a. Which tools have more capabilities in terms of cleaning the texts, performing Text Mining, and Sentiment Analysis?
b. Which tools perform better in terms of execution of code/node, results exportation, graphical capabilities and user-friendliness?
c. Which features, or functions are important in a Text Mining and Sentiment Analysis tool?

## 2.0 LITERATURE REVIEW

Text mining (TM) is also known as text data mining or text analytics. It is a procedure of extract, evaluate information and transform that information into valuable information for business benefits. [26] stated that TM is a tool that discover new hidden information within new or previous information by extracting information from various written resources. Besides, [27] stated that TM refers to the process of obtain interesting and relevant information and identify pattern or knowledge from the unstructured documents. In addition, according to [28], TM is defined as a process of exploring a huge collection of documents for discover and utilize the knowledge that found in the document. Also, in [29] mentioned that even though Natural Language Processing (NLP) and knowledge extraction are essential steps when processing TM through adding the value to data cleaning steps to change unstructured data into structured data that appropriate for complicated analysis. However, NLP and knowledge extraction that concern on summing up document independently within the collection unlike TM includes identifying patterns or trends that cover the whole collection of the document. Text Mining process consists of the following basic stages (see Figure 1):

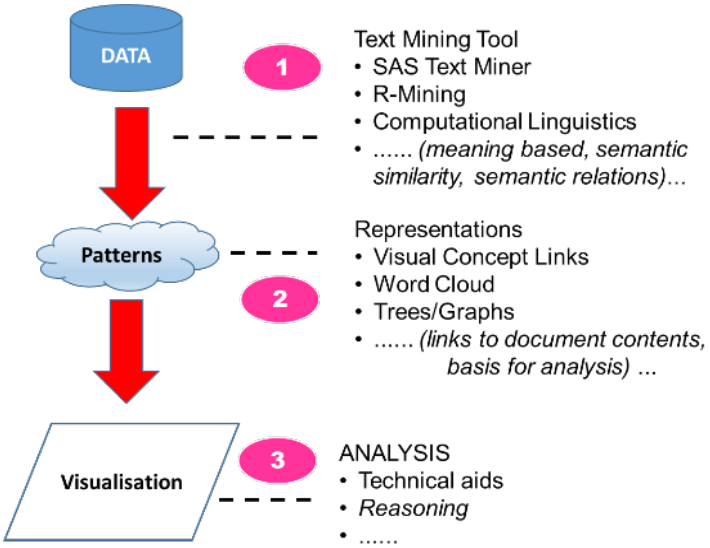


Figure 1. Text mining process adopted from [6]

Text Pre-processing or Text Parsing and Filtering - In this step, the text will be filter, parse and dictionary will be created according to machine language technique. Besides, text pre-processing is further categorized into 3 main phases:

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my 

a. Tokenization - The text will be segment into word and remove the blank space, comma, etc of the text document.
b. Stop word removal - In this step, the tags such as XML, HTML etc. from the webpages will be eliminated. The stop words such as “a”, “of” “the”, “is”, etc. will be removed after the process of removing tags.
c. Stemming - The process of identify the origin of specific word will be involving.

Text Transformation - In this step, the text will represent by the word it carries and the number the word appears in the text document. There are two methods to represent text document:

a. Bags of words
b. Vector space

Feature Selection - In this step, variable selection is involved. Process of selecting relevant or concern variables to use in building model.

TM Selection - In this step, method of data mining involved. Approaches such as clustering, predictive analysis, association classification etc. to extract the insight of the text document.

Interpretation / Evaluation - In this step, the result of the TM will be interpreted.

According to Jalanila & Subramanian [15], TM tools can be classified into three main categories:

**i.** Proprietary - These TM tools are under or owned by a company. Firms that need to analyse to use these TM tools are required to purchase. Even though these tools are available in free, but the feature of the tools are limited for the user to utilize. For example, SAS and Discover text.
**ii.** Open source - These tools are available for free, and user could download within the TM website. For instances, Python, R programming, and KNIME.
**iii.** Online - These online tools can be accessed from the website itself without downloading them. However, these tools only provide simple and limited functionality. For instance, text analyser, Voyant, etc.

## 3.0 METHODOLOGY

The method of research for the features, functions and key characteristics of the tools is conducted by reading product documentation online. In this study, SAS Enterprise Miner which has the text mining features [30], Python and R [31] were used to conduct Text Cleaning, Text Mining and Sentiment Analysis. As a result, the capabilities, features, and functions were documented into a summary table as the result of the comparative analysis.

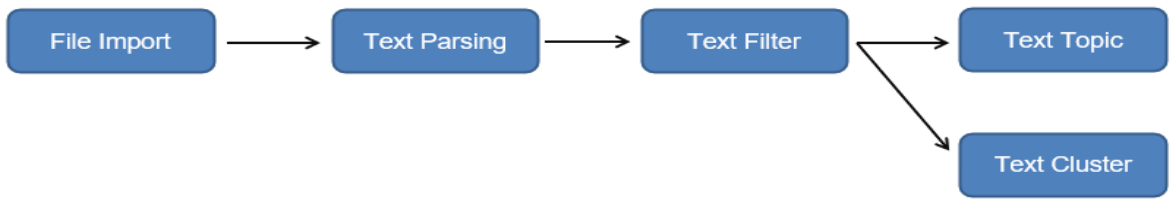


Figure 2. Process flow diagram in SAS Enterprise Miner

Figure 2 shows the process flow diagram for Text Cleaning, Text Mining and Sentiment Analysis, which is conducted in the SAS Enterprise Miner [20]. It consists of 5 main nodes: (i) File import, (ii) Text Parsing, (iii) Text Filter, (iv) Text Topic and (v) Text Cluster.

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my 


Figure 3. Process Flow diagram in Python and R

On the other hand, Figure 3 illustrates a process flow diagram for Text Cleaning, Text Mining and Sentiment Analysis which is commonly used in in Python and R. It comprises 4 common steps – (i) obtaining data (corpus), (ii) text processing, (iii) text mining and (iv) sentiment analysis. The evaluation is, these tools will be compared based on aspects such as the ability to perform parsing, stemming, filtering, checking and correcting spelling, removing stop-words (text pre-processing); to obtain term frequency and term association (text mining); to conduct Sentiment Analysis on Document-level, Sentence-level and Entity-/ Aspect- level and obtaining results for sentiment scores, segregation of positive and negative terms (sentiment analysis).

Their features, functions and key characteristics that are relevant to Text Mining and Sentiment Analysis were also discussed based on the data handling capabilities of the tools, execution of codes and nodes, exportation of results (output and graph), graphical capabilities, user-friendliness, customer support service and support, and community, ease of learning of the tools. Other aspects were included in the summary table to clarify the version of software used to conduct this research and the analysis. Suggestions were provided by proposing mandatory features and functions to be included in a Text Mining and Sentiment Analysis tool. Improvements based on insights gained on the shortcomings were also suggested after utilizing the tools. Recommendations to conduct Text Mining and Sentiment Analysis and considerations that should be taken into accounts during the analysis were also discussed.

## 4.0 RESULTS AND ANALYSIS

This sub section discusses a comparative analysis of SAS Enterprise Miner (proprietary text mining tool), Python and R programming (open-source text mining tools) respectively based on their capabilities.

### 4.1 Comparative Analysis Of SAS, Python, And R

Table 1 shows the comparative analysis of SAS, Python and R for text cleaning tasks. It is observed that SAS has strongest capabilities in performing all necessary tasks in cleaning the texts when compared to Python and R. Python has errors in performing stemming, and spelling checking and correcting, and is also unable to perform text parsing. R was unable to perform parsing as well and has errors while performing stemming.

Table 1. Text cleaning tasks using SAS, Python, and R

| Text Cleaning Tasks | SAS | Python | R |
|---|---|---|---|
| Remove Twitter username | ✓ | ✓ | ✓ |
| Remove URLs | ✓ | ✓ | ✓ |
| Remove Punctuations | ✓ | ✓ | ✓ |
| Remove Symbols | ✓ | ✓ | ✓ |
| Remove Numbers | ✓ | ✓ | ✓ |
| Remove Stop-words | ✓ | ✓ | ✓ |
| Perform Parsing | ✓ | X | X |
| Perform Stemming | ✓ | X | X |
| Spelling Checking and Correcting | ✓ | X | ✓ |
| Perform Text Filtering | ✓ | ✓ | ✓ |

### 4.2 Text Mining Capabilities

It was discovered that SAS and R enabled results output for term frequency and term association while python has problem in obtaining term association (see Table 2). This could be caused by the fact that

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my

Python just started venturing into Text Mining in the recent years hence the packages still have large room for improvements.

Table 2. Text Mining tasks using SAS, Python, and R

| Text Cleaning Tasks | SAS | Python | R |
|---|---|---|---|
| Obtaining Term Frequency | ✓ | ✓ | ✓ |
| Obtaining Term Association | ✓ | X | ✓ |

### 4.3 Sentiment Analysis Capabilities

Since there was no access to SAS Sentiment Analysis, only Entity-/ Aspect- level Sentiment Analysis could be conducted using SAS Enterprise Miner that has SAS Text Miner license. SAS Sentiment Analysis was required to conduct Sentiment Analysis on all levels. For Python, selected packages only enabled document-level Sentiment Analysis to be conducted to obtain sentiment scores while R has various packages to support Sentiment Analysis to be conducted on all levels (See Table 3).

Table 3: Sentiment Analysis using SAS, Python, and R

| Sentiment Analysis Levels | Sentiment Analysis Outputs | SAS | Python | R |
|---|---|---|---|---|
| Document-level | Sentiment Score | X | ✓ | ✓ |
| Sentence-level | Positive and Negative words | X | X | ✓ |
| Entity-/ Aspect- level | Insights derived from results | ✓ | X | ✓ |

### 4.4 Features And Functions Of SAS, Python, And R

It was found out that Python requires the longest execution time compared to SAS and R while the difference between SAS and R is not significant (see Table 4). The results exportation is easier for SAS as users only need to use the built-in drag and drop option, but codes are required to be scripted in order to export certain Text Mining and Sentiment Analysis results from Python and R. Python and R has greater graphical capabilities compared to SAS since Python and R can generate graphics with more flexibilities in terms of its' colour, size and design. This is enabled by the manipulation done in scripting codes. SAS has high user-friendliness as the drag and drop user interface and rich product document provide great support [6, 32, 33]. R has middle user-friendliness as scripting codes are still required to conduct Text Cleaning, Text Mining and Sentiment Analysis, this also means that users need to have extensive coding knowledge and expertise, but their product documentation (based on packages) is always available for users. RStudio has an interface that allows results to be in "all-in-one" view manner (Script window, Terminal window, Variable window and Graph window). Python has low user-friendliness as the tool only has a scripting window, the codes and results are in the same window, it is possible to get messy at times.

There is a lack of official product documentation since it is an open-source tool, users can only rely on the support from Python community. Users need time to get familiarized with the usability of the functions to match their research analysis, objectives and purposes.

Table 4: Features and functions of SAS, Python, and R

| Features and Functions | SAS | Python | R |
|---|---|---|---|
| Execution time of nodes/ codes | Excellent | Poor | Good |
| Results exportation | Easier (one-click) | Harder | Harder |
| Graphical capabilities | Fair | Good | Good |
| User-friendliness | High | Low | Middle |
| Overall Performance | Excellent | Poor | Good |

In general, SAS has the best performance because the tool has stronger text cleaning capabilities [33,35,36] compared to Python and R. Text Cleaning always precede Text Mining and Sentiment Analysis, hence stronger Text Cleaning capabilities can always yield better results for Text Mining and Sentiment Analysis. R has better performance compared to Python because even though both open-source tools have weaker text cleaning capabilities, R stood out in having a more developed packages developed to run comprehensive Text Mining and Sentiment Analysis. 'SentimentAnalysis' and 'sentimentr' packages are some of the powerful packages utilized to run Sentiment Analysis for this comparative study [13, 34].

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my 

## 5.0 DISCUSSION

After evaluating SAS, Python, and R on their capabilities in terms of Text Cleaning, Text Mining and Sentiment Analysis, it was suggested that the Text Cleaning capabilities of both open-source tools need to be diagnosed and improved to yield better Text Mining and Sentiment Analysis results. This is since the accuracy, reliability and consistency of results will be mandatorily affected by the output gained from Text Cleaning process, inevitably. Developers need to place great emphasis on improving the Natural Language Processing algorithm used to parse, stem and filter texts. It was clear that Python still has much more to catch up on, the algorithms used to manipulate texts and run analysis such as Text Mining and Sentiment Analysis need to be built in a more comprehensive way so that the results gained are with utmost usability and reliability. A comprehensive Text Mining and Sentiment Analysis tool should include a user-friendly interface and it should allow easy exportation of results, coding should be at minimal, and the tool should be supported by concise documentation and strong community. Features regarding Text Cleaning should be put on emphasis as Text Cleaning is a pre-requisite to Text Mining and Sentiment Analysis. Text Mining features that can perform text parsing, text stemming, text filtering, and spelling-checking and correcting should be included in the tool mandatorily because these features have direct impact on the consistencies of the text used for Text Mining and Sentiment Analysis.

For Text Mining, the tool should enable the generation of results regarding Term Frequency and Term Association. It was suggested that the tool mandatorily enable result generation for Term Association as this feature will indirectly facilitate Sentiment Analysis to be conducted on an Entity-/ Aspect- level that will help to drive useful insights. For Sentiment Analysis, the tool should enable Sentiment Analysis to be conducted on Document-level, Sentence-level and Entity-/ Aspect- level by enabling the generation of results regarding sentiment scores, the segregation of positive and negative terms and results generated based on keywords (Entity-/ Aspect- level based approach).

## 6.0 CONCLUSION

In terms of Text Cleaning, Text Mining and Sentiment Analysis, SAS, Python and R have their own capabilities that will affect each other sequentially and accordingly. To answer the research questions, a combination of SAS and R would yield better results of Text Mining and Sentiment Analysis as both tools complement each other's flaws in manipulating texts and running Sentiment Analysis on all-levels. Python is still a new player in the field of Text Mining and Sentiment Analysis, more developments need to be taken to increase results accuracy and usability. In this study the number of data being used are not important as the results of this study focused on functions and features of each tool. Future studies should be done in conducting different tasks relevant to Sentiment Analysis, and Text Mining and Sentiment Analysis tools should be improved to enable all levels of Sentiment Analysis. The Natural Language Processing algorithm should really be investigated the improvement of the open-source Text Mining and Sentiment Analysis tool. Proprietary Text Mining and Sentiment Analysis tool should also expand its offering by enabling all levels of Sentiment Analysis to be conducted. Other than those, new and available packages for Text Cleaning, Text Mining and Sentiment Analysis using Python and R should also be tested and documented to expand on the evaluation on this comparison study.

## 7.0 ACKNOWLEDGMENT

The authors would like to thank all the relevant parties in supporting this research works especially to Sunway University and National Defence University of Malaysia (NDUM) for supporting this research.

## List of Reference

[1] Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR), 49(2), 1-41.

[2] Zainol, Z., Jaymes, M. T., & Nohuddin, P. N. (2018, May). Visualurtext: a text analytics tool for unstructured textual data. In Journal of Physics: Conference Series (Vol. 1018, No. 1, p. 012011). IOP Publishing.

[3] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing: Proceeding of CISC 2017 (pp. 639-647). Springer Singapore.

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my 

[4] Ibrahim, N. F., & Wang, X. (2019). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. Computers in Human Behavior, 96, 32-45.
[5] Zainol, Z., Lee, A. S., Nohuddin, P. N., Ibrahim, N. F., & Hijazi, M. H. A. (2022). Examining the relationship of keyword analysis using online traveller hotel reviews. International Journal on Perceptive and Cognitive Computing, 8(1), 47-52.
[6] Lee, A. S., Yusoff, Z., Zainol, Z., & Pillai, V. (2018). Know your hotels well! An online review analysis using text analytics. International Journal of Engineering & Technology, 7(4.31), 341-347.
[7] Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. Expert Systems with Applications, 167, 114324.
[8] Chauhan, P., Sharma, N., & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. Journal of Ambient Intelligence and Humanized Computing, 12, 2601-2627.
[9] Zainol, Z., Nohuddin, P. N., Lee, A. S. H., Ibrahim, N. F., Yee, L. H., & Majid, K. A. (2021). Analysing political candidates' popularity on social media using POPularity MONitoring (POPMON). SEARCH Journal of Media and Communication Research, 39-55.
[10] Zainol, Z., Wani, S., Nohuddin, P. N., Noormanshah, W. M., & Marzukhi, S. (2018). Association analysis of cyberbullying on social media using Apriori algorithm. International Journal of Engineering & Technology, 7(4.29), 72-75.
[11] Chelmis, C., Zois, D. S., & Yao, M. (2017, November). Mining patterns of cyberbullying on twitter. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 126-133). IEEE.
[12] Rezvani, N., & Beheshti, A. (2021). Towards attention-based context-boosted cyberbullying detection in social media. J. Data Intell, 2, 418-433.
[13] Lee, A., Lim, T., Chia, M., Ea, L., & Yap, M. Y. (2017). Mining "What they talk about" for a private healthcare service provider. Archives of Business Research, 5(5), 135-156.
[14] Semwal, T., Milton, K., Jepson, R., & Kelly, M. P. (2021). Tweeting about twenty: an analysis of interest, public sentiments and opinion about 20mph speed restrictions in two UK cities. BMC public health, 21(1), 2016.
[15] Jalanila, A., & Subramanian, N. (2016, October). Comparing SAS® Text Miner, Python, R: analysis on random forest and SVM models for text mining. In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 316-316). IEEE Computer Society.
[16] Kaur, A., & Chopra, D. (2016, September). Comparison of text mining tools. In 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 186-192). IEEE.
[17] Lawrence, L. (2014). Reliability of sentiment mining tools: a comparison of semantria and social mention (Bachelor's thesis, University of Twente).
[18] Gandharv, S., Richhariya, V., & Richhariya, V. (2017). Real time text mining on twitter data. International Journal of Computer Applications, 178(3), 24-28.
[19] Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N., & Topi, H. (2003). A review of software packages for data mining. The American Statistician, 57(4), 290-309.
[20] Vyas, V., & Uma, V. J. P. C. S. (2018). An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. Procedia Computer Science, 125, 329-335.
[21] Stavrakantonakis, I., Gagiu, A. E., Kasper, H., Toma, I., & Thalhammer, A. (2012). An approach for evaluation of social media monitoring tools. Common Value Management, 52(1), 52-64.
[22] Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).
[23] Zhang, Q., & Segall, R. S. (2010). Review of data, text and web mining software. Kybernetes, 39(4), 625-655.
[24] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2), 1-33.
[25] Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18-38.
[26] Petrou, C. (2005). Text mining: A tool for statistical learning. University of Louisville.
[27] Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases (Vol. 8, pp. 65-70).
[28] Haider, M., & Gandomi, A. (2021). When big data made the headlines: Mining the text of big data coverage in the news media. International Journal of Services Technology and Management, 27(1-2), 23-50.
[29] Advantage, C. B. SAS® Text Miner.

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my 

[30] Barańska, K., Różańska, A., Maćkowska, S., Rojewska, K., & Spinczyk, D. (2022). Determining the intensity of basic emotions among people suffering from anorexia nervosa based on free statements about their body. Electronics, 11(1), 138.
[31] Yang, D. (2018). SolarData: An R package for easy access of publicly available solar datasets. Solar Energy, 171, A3-A12.
[32] Nicholas, C. K. W., & Lee, A. S. H. (2017, October). Voice of customers: Text analysis of hotel customer reviews (cleanliness, overall environment & value for money). In Proceedings of the 1st International Conference on Big Data Research (pp. 104-111).
[33] Lee, A. S. H., Daniel Chong, K. L., & Khin Whai, N. C. (2019). OpinionSeer: Text visualization on hotel customer reviews of services and physical environment. In Information science and applications 2018: ICISA 2018 (pp. 337-349). Springer Singapore.
[34] Lee, A., & Lim, T. M. (2016). Mining opinions from university students' feedback using text analytics. Information Technology in Industry, 4(1).

*Corresponding Author | Yi, F. L. | angelal@sunway.edu.my 