Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 3, June 2020, pp. 1303~1310
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i3.pp1303-1310

Text analysis on health product reviews using r approach

Nasibah Husna Mohd Kadir, Sharifah Aliman
Faculty of Computer and Mathematical Sciences, Advanced Analytics Engineering Center, Malaysia

Article Info

Article history:

Received Oct 7, 2019
Revised Dec 9, 2019
Accepted Dec 23, 2019

Keywords:

Big data
R programming
Text analysis
Unstructured data

ABSTRACT

In the social media, product reviews contain of text, emoticon, numbers and symbols that hard to identify the text summarization. Text analytics is one of the key techniques in exploring the unstructured data. The purpose of this study is solving the unstructured data by sort and summarizes the review data through a Web-Based Text Analytics using R approach. According to the comparative table between studies in Natural Language Processing (NLP) features, it was observed that Web-Based Text Analytics using R approach can analyze the unstructured data by using the data processing package in R. It combines all the NLP features in the menu part of the text analytics process in steps and it is labeled to make it easier for users to view all the text summarization. This study uses health product review from Shaklee as the data set. The proposed approach shows the acceptable performance in terms of system features execution compared with the baseline model system.

Corresponding Author:

Sharifah Aliman,
Faculty of Computer and Mathematical Sciences,
Advanced Analytics Engineering Center,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.
Email: sharifahali@tmsk.uitm.edu.my

1. INTRODUCTION

Product review plays an important role in the consumer purchase decision and the manufacturer business insights [1]. The manufacturers tend to have the insights from consumer reviews of the products, to help in generating ideas for marketing strategies to expand the products [2]. People in various businesses can derive useful information from social network data to understand their consumers more comprehensively and precisely by utilizing various types of social media analytic tools [3]. In reaching the variation of social media data, consumers products review in social media are in unstructured form [4-6]. The review contains of text, emoticon, numbers and symbols that hard to identify the text summarization which is harder to sort and categorize them into the useful summaries [7]. This study contribution covering the R approach of text analytics using web-based in the health products consumers online review in capturing insights and view from social media data for the health products. The lack of web interface in the analytics area makes marketer and people in business world cannot have the data translation from the big data.

Furthermore, from various sources of social media [8], a business will traditionally look from one to one platform to get feedback on their products [9-10]. Distribution and split of this application also cause efficiency interruption in generating information from consumer review data [11]. Many analysts can do analytics products to determine the details of data intent but lack of integration between data translation with the web interface. In analyzing the unstructured textual data, Batrinca et al [12] categorized the key techniques into six that are natural language processing (NLP) [13], news analytics, opinion mining [14-16], data scraping, sentiment analysis [9, 17-23] and text analytics [10, 24]. In this study, initially, natural language processing is the procedure of strategy to separating significant data from regular language input and creating tendency data meaning. The second technique, text analytics is used to distinguish and extricate emotional data in source materials. These two techniques are relying each other because these elements use the same features in obtaining text or sentiment summarization. This study consists of Section 2 that explains

the related works from the comparison table of the NLP studies, Section 3 that explains the proposed system R approach, Section 4 discuss the results of the approach and Section 5 contain the conclusions.

2. RELATED WORK

There are similar applications related to text analytics method. For the data used by these similar applications are from social media sources such as Facebook [25], Twitter [26, 27] and there are some resources taken from a business-theme website such as online review data from hotel booking websites [24] and online product sales. From the comparative studies made all of the data taken from online sources and requires the process of identifying the required data and data forms. This is because data obtained from online sources are not the same according to the type and form of the database used by the online source platform. Therefore, each similar application using different tools and methods in obtaining data extracted from the online source platform. Enlightenment for this extraction data is important as the information required by the manufacturer depends on the frequency of data and the number of responses from customers to their product or business.

Referring to the related works, there are two types of Natural Language Processing used by similar applications previously called Sentiment Analysis and Text Analytics. For sentiment analysis, it was carried out to find out the situation and tendency in a situation such as the example in the study of Al-Saffar, et al. [18] which studies the Malay Text Model Classification using sentiment analysis. Next, et al. [19] and Kamyab, et al. [17] have used the sentiment analysis method in their study in identifying social sentiment through social media, using Twitter post data and the country's website against the country and political issues. This suggests that, sentiment analysis and text analytics are among the ways to process texts and opinions that are voiced by people through social media mediums. This method is widely used because through research that has been made, areas that use sentiment analysis and text analytics include healthcare, politics, country and business. These are the major areas that involve direct contact with humans. From the obtained study, it also shows that in 2018 various studies were conducted on sentiment analysis. While for text analytics is still low.

Through this observation, it can be concluded that the above studies do not have user interfaces that may be due to a non-focused study objective of the user interface or a limited research capability on the generation of this user interface. In addition, all of the above studies have features that are used in displaying analytical results from sentiment and text analysis works that have been done. As observed, the extraction and term frequency data process is a mandatory process that has been done by all of the above studies. Next are the features analysis results used by these studies are Word Cloud and Bar Plot. These two features are graphical data summarization to the sentiments and text analytics performed by the studies. Understanding on what's behind this matter, sentiment and text analysis that can give insights should be further expanded to gain the breadth of generating information and the quality of an insight.

3. PROPOSED FRAMEWORK WEB-BASED TEXT ANALYTICS

The main objective of this study is to solve the unstructured data by sort and summarize the review data using a Web-Based Text Analytics using R approach. The process of text analytics was including scraping data on the social media using schematic code and tools, association analysis and visualization in order to predictive analytics. As the nature and type of social data represent the unique attributes of big data in aspects of volume, variety, velocity, veracity [4]. Figure 1 shows the proposed framework of web-based text analytics for the study.

The first phase objective is to extract the online consumer reviews from social media related to Shaklee health products reviews. In identifying the related health product and the social media platform, we perform the comparative table of similar application about text analytics and the potential system design. The data extraction tool is the important element to extract the data review from the social media. In retrieving the social media product review data, system developer needs to have permission form the social media owner except the open public social media. Some social media like Twitter and Facebook need to approve the permission of entering their data system. The API permission from social media owner consists of keys and access token secret for researcher accessing the social media data. The extracted data then is transformed into the CSV file format as the dataset will be used in the text analytics system. A web has a huge amount of data and the web also developed using a combination of different programming or depending on the developer of the web. Therefore, to get the desired data from a web, web scraping is a technique for extracting data from a web site. Web scraping will extract unstructured or semi-structured data into a database that has an organized structure because the structured intent here is structured.

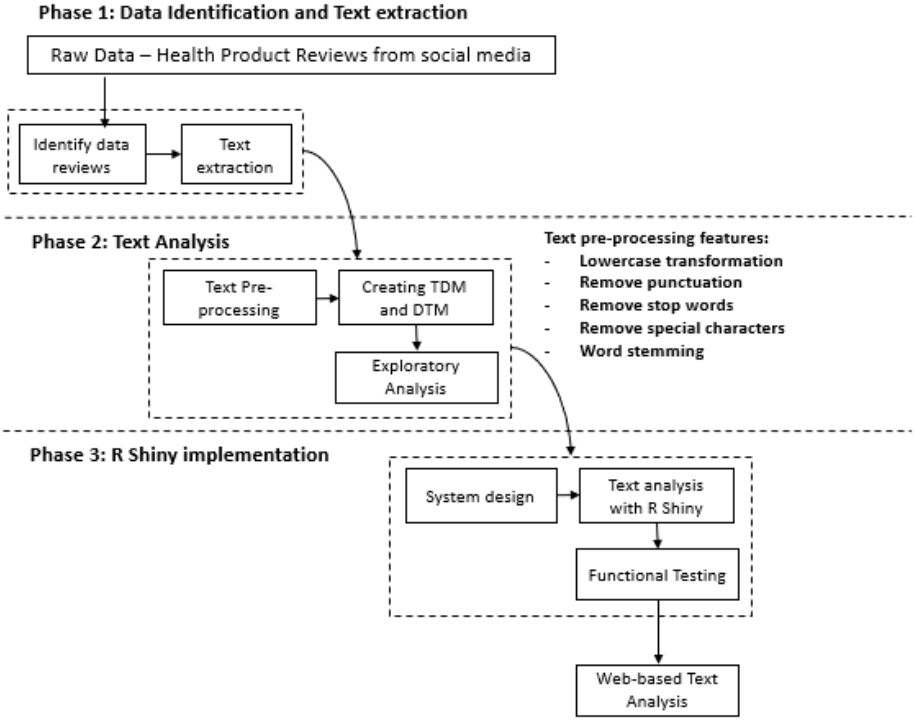

Figure 1. Proposed framework web-based text analytics

Next, the second phase objective is to analyze the online consumer reviews using text analytics by using R packages approach. R packages includes the text pre-processing method to perform lowercase transformation, remove punctuation, remove stop words, remove special characters and word stemming. Text pre-processing performs the review transformation in order to standardize the output of the text review before the text been analyze and processed. According to [8] there are four process related to text analysis work that are importing the dataset, cleaning and pre-processing the raw data, representing, filtering and weighting the tokens and lastly is analyzing. First step starts with importing the many types of text for example in .pdf, .csv or excel files into the R text corpus using the R package for running the operations. Second technique is cleaning and pre-processing the documents. In this step, it covers the stemmization, remove stop words, punctuation and transform the string into the lower case and then this string become into tokens. Then, the third procedure is representing, filtering and weighting the tokens into DTM (Document Term Matrix), Token list and TF-IDF in representing the tokens corpus of the document files datasets. Lastly, is analyzing the DTM and Token list into the visualization of graph or bar plot or in the summarization of words of word cloud. At the last process, the text summarization will result the data into four (4) features that are word cloud, word tokenizer, word breakdown and word count bar plot by using the R Shiny application.

The last phase objective is developing prototype web-based of text analysis of health product reviews. The R Shiny was implemented in the R Studio to perform the web-based system. R Shiny is the packages that combining the element of upload data modules, view data modules, word breakdown function, word tokenizer function, word cloud function and bar plot function. R Shiny as the tools to perform the data into web-based system and have a wide of R packages to supporting this analytics project. R. Shiny connects between codes and interfaces that help facilitate product owners to understand the continuously charging of big data reviews about their products. The process for using R Shiny is that we install R Shiny package into R using ***'install.package (shiny)'*** code, next to use R Shiny use library code (shiny) to insert the package into the library. Selection of R Shiny is very effective because R provides a lot of libraries not only to do text analysis process, but also to load all analysis processes into the interface and the web.

R is open source [10] and the environment for statistical computing or statistical programming language and graphic that is frequently been used by statisticians and professionals from many fields to perform data analysis [28, 29] One of the main purposes for R is to group the data set of product review into data classification [30]. As stated in this research project [31], there are 594 R packages records that are used for a variety of uses i.e. solving business problems related to key areas such as Retail, Health, Insurance and Politics [14].

Referring to the study in [17-19], all of these projects use R as a tool to analyze the data obtained according to their respective field of study. Generally, all data processing uses Data Process Extraction, Term Frequency for TDM-DTM and for two features are Word Cloud and Bar Plot. Inside R, the packages used to carry out these studies are:

R Packages for not web-based Data Analysis in R, [31]:

a) 'tm' – for text mining
b) 'SnowballC' – for text stemming
c) 'wordcloud' – for word cloud generator
d) 'RColorBrewer' – for color palettes

4. RESULTS AND ANALYSIS OF EXPERIMENTS

This study used the R as the programming tools, R Shiny as the tools to perform the data into web-based system and have a wide of R packages to supporting this analytics project. For this study, we executed the prototype using Windows 10 operating system, Intel core-i5 and 4GB installed RAM. We used Shaklee [32] review data as the health product dataset. The dataset obtained from the *Cari Forum* website [27, 33]. In our dataset there are 2,515 of online message reviews we extracted from the *Cari Forum* using the *Import.io* tools.

4.1. Web-Based Interface

Since web-based interfaces are the intermediaries of the relationship between data and humans via the web, this study has developed the front-end and back-end interface. The front-end will be used by users who manipulate the data, while back-end is an infrastructure that supports all movement of data that will be used by end users [34].

Figure 2 shows that the back-end interface is important to inform users about how online data is extracted. The extracted data is online streaming data which requires certain tools to realize the alignment between the system and the data that is to be obtained. This means that the data we want to get is coming from the source of the web or the unknown page of the infrastructure and the system behind the web or the page. Therefore, the tools used in the project enable the data to be read on the surface only which does not involve the merger or consolidation of any system of systems. The conclusion is that the data extraction process is faster and more efficient by using tools because the user can select the part of the review data that is to be used only that is related to the product and the data obtained is in much quantity.

Figure 3 shows the front-end interface for uploading data. This interface permits the user to upload the online review data which in the TXT or CSV file format into the system. The user then clicks on the upload data button to perform the file upload process. The file must in the .txt or .csv file format and the text file is a data input to the system process. Upload data is in the second phase of framework because the upload data will be processed using the back-end process.

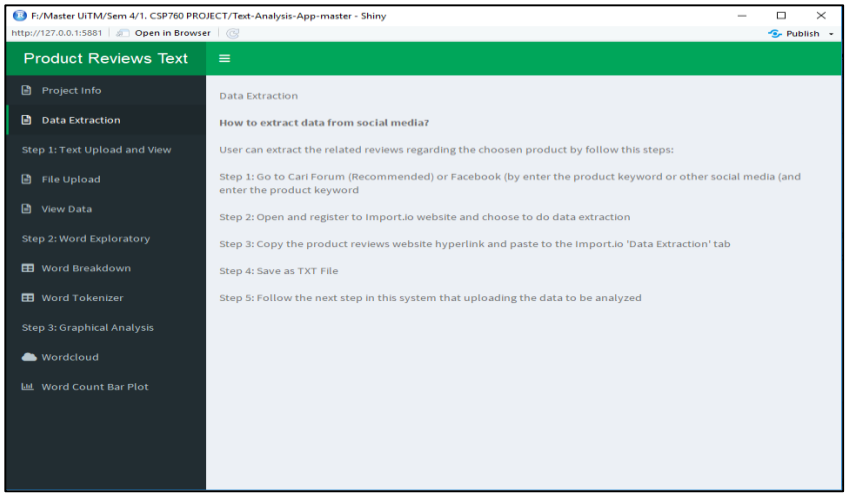

Figure 2. Data extraction introduction interface

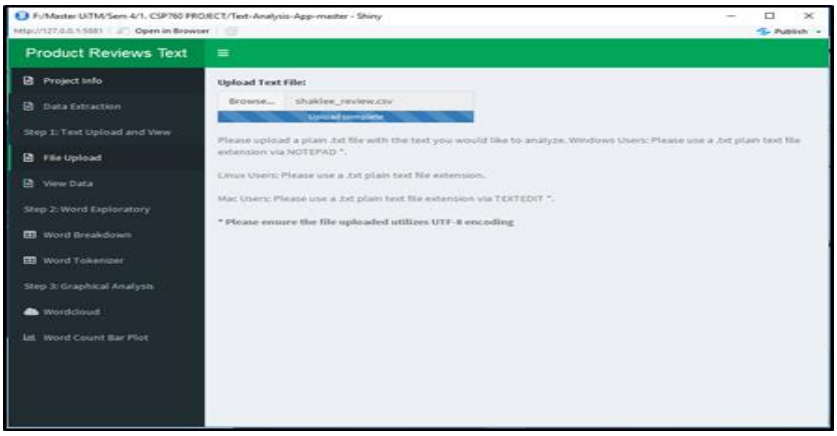

Figure 3. File upload interface

After the data have been uploading in the system, the system permits the user to view the data upload. In the web-based system also shows the inserted data that can be view by user using this module. The review data that has been inserted into the database has been displayed in the viewing space of those reviews. Overall review data will be displayed, and the data is connected to each other. The review data is the original review data that has stop words, symbols, special characters and numbers that have not been removed and will be discarded during the pre-processing process to run the next process of text analysis. In conclusion, this text preview is very important for us to see the summary and combinations of reviews on the product roughly further make the difference before and after the data are analyzed.

4.2. Exploratory Analysis and R Shiny Implementation

Exploratory analysis is a critical process to investigate data to see patterns of data, to detect anomalies, to conduct hypothesis testing and to examine the budget with the help of summary statistics and graphical representations [35]. Therefore, summary statistics and graphical representation in this project are four Word Breakdown, Word Tokenizer, Word Cloud and Bar plot. For summary statistics, there are two features, Word Breakdown and Word Tokenizer, while for graphical representation are Word Cloud and Bar plot. For each feature, it has its own uses in helping to gain insights to users.

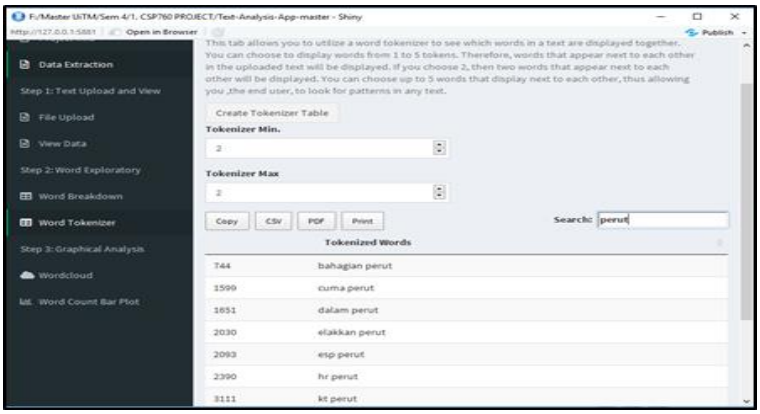

Figure 4. Word tokenizer interface

Figure 4 shows the second exploratory analysis of text analytics which is the back-end interface, Word Tokenizer Interface. Before this word tokenizer is generated, the system called the data parameter to perform this analysis. Word tokenizer is the process to see the words that come out with a certain word. As the example shown in Figure 4 when we select the word 'belly', the automatic system will display all entries from the databases that are related to the 'stomach' i.e. their tokenized words are 'the abdomen,' 'Just stomach' and other entries that have the term in it. From this view, this Word Tokenizer function is important in generates idea and insight about the frequency term of words and the words come with that thrown by

consumer according to the products. In addition to the main function of removing the Word Tokenizer view, this page also has the option of Tokenizer min and max for us to select the token or word value that we want to see and study.

Figure 5 shows another back-end interface called Word Cloud interface of text analytics. This interface shows the system can generate the word cloud based on the term frequency of the online data review. Word Cloud will show the user the frequent words incoming from the consumer about the products. The outgoing word will be arranged as cloud, which has great value varying by the frequency value of the word. From this word cloud analysis, it provides a holistic view of the word that comes about the product. Users can see clearly the words used by consumers that may consist of words about product value, product quality and satisfaction when using the product.

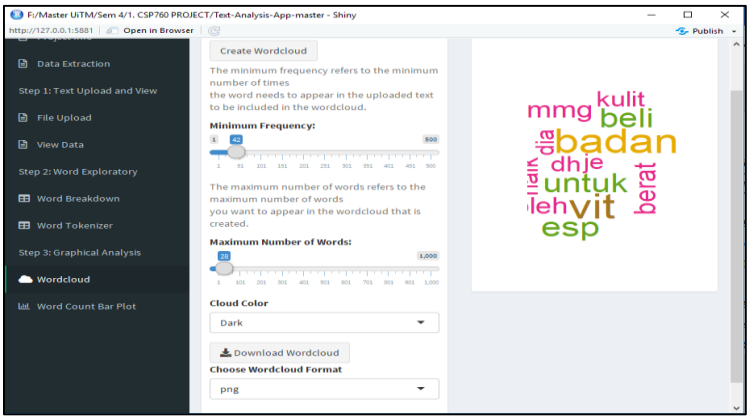

Figure 5. Word cloud Interface

4.2.1 Graphical Analysis: Word Cloud

The graphical results for text analysis are shown in Word Cloud form. Word Cloud will collect all terms with frequency terms from high to low. Next it will be shown in the form of clouds having the difference in the size of the writing according to the frequency terms which is large for the high frequency value, while the smaller for the low frequency value.

There were 12 terms retrieved (see Figure 6) from 2 pages generated through the Word Cloud analysis process. The words collected were Shaklee health products related terms such as the product categories, products effectiveness, and involvement of members in Shaklee activities. Through this result, we can determine the importance of a product. From the above example, the product names are ESP and Alfalfa, both of which are Shaklee products that are highly termed in the review data. This means that these two products may be Shaklee's favorite product or hot key product as the ability of the product to help treat various diseases or in balancing nutrients in the body. As such, this company can increase the marketing activities of other products too for consumers to see and consume.

“vit”,”vitamin”,”esp”,”berat”,”muka”,”kulit”,”badan”,”alfalfa”,”membantu”,”pengguna”, ”beli”,”harga”

Figure 6. List of word terms to be generate

4.2.2 Graphical Analysis: Bar Chart

Inside Web-based, the bar chart provides a smoother visual and structured in providing an overview through the bar chart. The resulting bar chart displays the result of frequency terms via the bar. The highest bar represents the term that has the highest, as well as for the medium and the lowest represents the low term frequency.

Figure 7 shows the term’s frequency of the highest frequency and sequential to low frequency terms. The chart showed 10 terms - *vitamin, vits, badan, esp, beli, boleh, berat, mmg, kulit* and *naik*. The highest word frequency is *vitamin* and the lowest is *kulit*. According to SME, most of words give the business insight

such as vitamin, badan, esp, beli, berat, kulit and naik. These terms show about the products effects to consumers from many parts of perspectives. This bar chart is a comprehensive overview that does not put specific values on each bar. However, the highlighted term is helpful in analysts to assess the level of usage and the frequency of terms mentioned in the review data of the product.

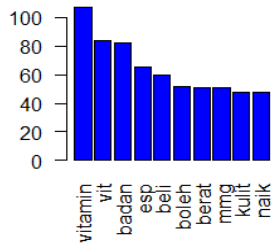

Figure 7. Graphical analysis of bar chart in web-based

5. CONCLUSION

As a conclusion, the used of R libraries and its text analytics packages in developing the web-based prototype is quite promising. The study has done the integration of front-end and back-end interface for web-based text analytics using R platform gives transparent text analysis at each module. At each module, user used the front-end interface to look at the implementation results without looking at the back-end processes such as stemming, generating frequency and generating suitable charts for visualization. The results of word cloud for Shaklee health products reviews might help Shaklee users to look what are best products in market and Shaklee outlets to strengthen their business and marketing strategies.

ACKNOWLEDGEMENTS

The authors would like to thank Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, for sponsoring this research.

REFERENCES

[1] Li, M., *et al*., (2013). "Helpfulness of online product reviews as seen by consumers: Source and content features", *International Journal of Electronic Commerce*, 17(4), 101-136.
[2] Alzahrani, H. (2016). "Social Media Analytics using Data Mining". *Global Journal of Computer Science and Technology*.
[3] Greene, J. A., *et al*., (2011). "Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook", *Journal of general internal medicine*, 26(3), 287-292.
[4] Gandomi, A., *et al*., (2015). "Beyond the hype: Big data concepts, methods, and analytics", *International journal of information management*, 137-144.
[5] Fan, W., *et al.*, (2014). "The power of social media analytics". *Communication Acm*, 57(6), 74-81.
[6] Bhatt, A., *et al*., (2015). "Amazon Review Classification and Sentiment Analysis", *International Journal of Computer Science and Information Technologies*, 6(6), 5107-5110.
[7] Tundjungsari, V. (2013). "Business Intelligence with Social Media and Data Mining to Support Customer Satisfaction in Telecommunication Industry", *International Journal of Computer Science and Electronics Engineering (IJCSEE)*, Volume, 1
[8] Welbers, K., *et al*., (2017). "Text analysis in R", *Communication Methods and Measures*, 1(4), 245-265.
[9] Bhatt, A., *et al*., (2015). "Amazon Review Classification and Sentiment Analysis", *International Journal of Computer Science and Information Technologies*, 6(6), 5107-5110.
[10] Ruan, G., *et al*., (2014). "*Textrweb: Large-scale text analytics with r on the web*", Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment, p. 63.
[11] Dolgobrod, M. (30 May, 2013). Semantic Scholar. Retrieved from Semantic Scholar: https://pdfs.semanticscholar.org/308a/92b8e2bc03f855dc76f805c981af3d061efc.pdf
[12] Batrinca, B., *et al*., (2015). "Social media analytics: a survey of techniques, tools and platforms", *Ai & Society*, 89-116.
[13] S. Muthukumaran, *et al.*, (2017). "Text Analysis for Product Reviews for Sentiment Analysis using NLP Methods", *International Journal of Engineering Trends and Technology (IJEET)*, 474-480
[14] Islam, M., *et al*., (2011). "A systematic review on healthcare analytics: Application and theoretical perspective of data mining", *Healthcare*, Vol. 6, No. 2, p. 54 Multidisciplinary Digital Publishing Institute.

[15] Varangaonkar, A. (18 December, 2017). "9 Useful R Packages for NLP & Text Mining". Retrieved from: https://hub.packtpub.com/9-useful-r-packages-for-nlp-text-mining/
[16] P. Khanna, *et al.*, (2017). "Sentiment Analysis: An Approach to Opinion Mining from Twitter Data using R". *International Journal of Advanced Research in Computer Science*, 8(8), 252-256
[17] Kamyab, M., *et al*., (2018). "*Sentiment Analysis on Twitter: A text Mining Approach to the Afghanistan Status Reviews*". Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality, 14-19.
[18] Al-Saffar, *et al.*, M. (2018). "Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm", *PloS one*, 13(4).
[19] Chiu, S. I., *et al*., (2018). "*Predicting Political Tendency of Posts on Facebook*". Proceedings of the 2018 7th International C,onference on Software and Computer Applications, 110-114.
[20] Kontopoulos, E., *et al.*, (2013). "Ontology-based sentiment analysis of twitter posts". *Expert systems with applications*, 40(10), 4065-4074.
[21] Fikri, M., *et al*., (2019). "A Comparative Study of Sentiment Analysis using SVM and SentiWordNet". *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, 13(3), 902-909.
[22] Safrin, R., *et al*., (2017). "Sentiment analysis on online product review". *Int. Res. J. Engineering. Technology*, 4(04).
[23] Khanna, P., *et al*., (2017). "Sentiment analysis: an approach to opinion mining from Twitter data using R". *International Journal of Advanced Research in Computer Science*, 8(8), 1-5.
[24] Ting, P. J. L., *et al.*, (2017). "Using big data and text analytics to understand how customer experiences posted on yelp. com impact the hospitality industry". *Contemporary Management Research*, 13(2)
[25] "Facebook" [Online], Available from: https://facebook.com
[26] "Twitter" [Online], Available from: https://twitter.com
[27] Laksana, J., *et al.*, (2014). "*Indonesian Twitter text authority classification for government in Bandung*". In 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA) (pp. 129-134). IEEE.
[28] Kohli, S., *et al.*, (2014). "*Data analysis with R*". Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, 537-538.
[29] Varangaonkar, A. (18 December, 2017). "9 Useful R Packages for NLP & Text Mining". Retrieved from: https://hub.packtpub.com/9-useful-r-packages-for-nlp-text-mining/
[30] Projects, C.-r. (17 June, 2019). "CRAN Packages". Retrieved from CRAN:https://cran.r-project.org/web/packages/available_packages_by_name.html
[31] STHDA. (5 May, 2019). "Text Mining and Word Cloud Fundamental". Retrieved from STHDA: www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know
[32] "Cari Forum" [Online], Available from: https://mforum.cari.com.my.
[33] "Shaklee" [Online], Available from: https://www.shaklee.com.my/
[34] Mehren, R. (13 January, 2017). "Creating Web Interface". Retrieved from MakeUseOf: https://www.makeuseof.com/tag/creating-web-interfaces-start/
[35] Patil, P. (23 March, 2018). "What is Exploratory Data Analysis?" Retrieved from Medium: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
[36] S. Sangam, *et al.*, (2019). "Sentiment Classification of Social Media Reviews using Ensemble Classifier". *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, 16(1), 355-363

BIOGRAPHIES OF AUTHORS

Nasibah Husna Mohd Kadir is a Master of Computer Science (Web Technology) at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.

Sharifah Aliman is a computer science senior lecturer whose interests are in social computing, information and society, social media analytics as well as scientific and Information visualization. She is an active member of two research groups: Advanced Analytics Engineering Center (AAEC) and Interactive Computing & Communication Technology (ICCT). Currently, she is Postgraduate Research Coordinator at Faculty of Computer and Mathematical Sciences, UiTM Shah Alam. Her professional memberships are Malaysia Board of Technologists (2018-2019), IEEE Member, IEEE Computer Society Member and Microsoft Certificate Application Development (2007-2008). She received her PhD in IT and Quantitative Sciences (2017) from UiTM, MSc. in IT (2001) from Universiti Putra Malaysia and BSc. in Computer Science and minor in Mathematics (1991), Midwestern State University, Texas, USA. She continues training teachers as she got the certified trainer for Computational Thinking and Computer Science Teaching in 2017 by Malaysia Digital Economy Corporation(MDEC).