Sains Malaysiana 50(6)(2021): 1815-1825
http://doi.org/10.17576/jsm-2021-5006-26

A Bibliometric Analysis of COVID-19 Research in Malaysia using Latent Dirichlet Allocation

(Suatu Analisis Bibliometrik Kajian COVID-19 di Malaysia menggunakan Agihan Dirichlet Terpendam)

ZAMIRA HASANAH ZAMZURI*

ABSTRACT

Coronavirus COVID-19 shocking the whole world due to its highly contagious characteristics implicating not only public health, but also economy and social life. Since the effects are momentous, plenty of research have been conducted and still ongoing in order to study and to learn more about this virus and how it changing our daily life. In this paper, we explore 134 articles published in 2020 related to COVID-19 and narrowing the scope of study to Malaysia. An alternative route was taken by employing Latent Dirichlet Allocation (LDA) to identify underlying themes or topics in these publications. Two separate analyses were conducted, one is to the paper's titles and another one to the journal's names. The findings identified three topics for paper's titles data are clinical study, impact of COVID-19 on various fields and Movement Control Order (MCO). The last topic shows the locality criterion in the studied papers as the term MCO was only used in Malaysia. For the journal's names, three topics identified were medical study, public health also business and education. Two papers with the most number of citations are both in social sciences. Investigating the properties of these topics, we found that papers on clinical studies are the ones with more chance to be cited and published by reputable publishers. These findings may help researchers on planning and strategizing for future research on COVID-19 specifying on Malaysia cases.

Keywords: COVID-19; movement control order; social sciences

ABSTRAK

Koronavirus COVID-19 telah mengejutkan seluruh dunia kerana ciri penularan jangkitannya yang melibatkan bukan sahaja kesihatan awam, tetapi juga ekonomi dan kehidupan sosial. Oleh kerana kesannya sangat penting, banyak kajian telah dijalankan dan masih dijalankan untuk mengkaji dan mengetahui lebih lanjut tentang virus ini dan bagaimana ia mengubah kehidupan seharian kita. Dalam makalah ini, kami mengkaji 134 artikel yang diterbitkan pada tahun 2020 berkaitan dengan COVID-19 dan mengecilkan skop kajian kepada Malaysia. Satu langkah alternatif telah diambil dengan menggunakan Latent Dirichlet Allocation (LDA) untuk mengenal pasti tema atau topik yang mendasari penerbitan ini. Dua analisis berasingan telah dijalankan, satu kepada judul makalah dan satu lagi untuk nama jurnal. Hasil kajian telah mengenal pasti tiga topik untuk data tajuk kertas iaitu kajian klinikal, kesan COVID-19 dalam pelbagai bidang dan Perintah Kawalan Pergerakan (MCO). Topik terakhir menunjukkan kriteria lokaliti dalam makalah yang dikaji kerana istilah MCO hanya digunakan di Malaysia. Untuk nama jurnal, tiga topik yang dikenal pasti adalah kajian perubatan, kesihatan awam serta perniagaan dan pembelajaran. Dua makalah dengan petikan tertinggi adalah dalam bidang sains sosial. Dalam kajian sifat topik ini, kami mendapati bahawa makalah mengenai kajian klinikal mempunyai peluang lebih baik untuk dipetik dan diterbitkan oleh penerbit terkemuka. Penemuan ini dapat membantu para penyelidik untuk merancang dan menyusun strategi untuk penyelidikan masa depan mengenai COVID-19 khusus untuk kes-kes di Malaysia.

Kata kunci: COVID-19; perintah kawalan pergerakan; sains sosial

INTRODUCTION

Coronavirus disease (COVID-19) hits Wuhan, China in November 2019. Beginning from there, the spread of the virus has been all over the world, and verified as a pandemic by World Health Organization (WHO) in 2020. The first case recorded in Malaysia is by Chinese tourists on the 25th January 2020 in which 3 Chinese nationals who previously had close contact with an infected person in

Singapore (Elengoe 2020). However, the number of cases rose significantly due to Sri Petaling Tabligh gathering on 27th Feb to 1st March 2020 that implicates the government to issue the Movement Control Order (MCO) as a preventive measure. The MCO starts on 18th Mar 2020 and the number of cases continue to rise and reached its peak on April 2020. When the number of active cases slowly declined, the lockdown restrictions have been relaxed over the next several months. Since mid-September 2020, an outbreak of cases in Sabah, Selangor, Kuala Lumpur, Negeri Sembilan, Johor, Penang and Kedah led to a resurgence of COVID-19 cases throughout the country. Currently, Malaysia has issued the second version of MCO (MCOv2) starting from 12th Jan 2021 in Selangor, Sabah, Melaka, Johor, Penang and Kuala Lumpur. All states in Malaysia except Sarawak are now instructed to follow MCOv2 that is scheduled to end on 4th Feb 2021.

Since the COVID-19 is considered as a pandemic, the whole world was affected not only on health perspective, but also economy, social and a country as a whole. Researchers all over the world are conducting studies on their field related to COVID-19 situation. Special issues on Covid-19 were observed in many journals from various fields not only from health or medical.

A bibliometric analysis referring to an analysis on the academic papers to search for trend and pattern. Fan et al. (2020) compare between English and Chinese studies on COVID-19 and found that large portion of earlier publications are authored by Chinese researchers. The trend then changes when COVID-19 receives global attention, with more publications were written in English. Chahrour et al. (2020) explored the PubMed and WHO databases for COVID-19 related publications and looked into the distribution based on the countries and their characteristics. Aristovnik et al. (2020) compares 16,866 COVID-19 publications between science and social science. The empirical results show the domination of health sciences in terms of number of relevant publications and total citations, while physical sciences and social sciences and humanities lag behind ominously. Deng et al. (2020) found that only 4.1% of the publications related to COVID-19 have more than 100 citations. Verma and Gustafsson (2020) focusing on COVID-19 publications related to business fields.

From the collections of past research on the bibliometric analysis of COVID-19, none have been performed narrowing down to Malaysia scope. Hence, in this paper, we explore COVID-19 publications that focusing on Malaysia scenario or include Malaysia as comparison; narrowed down by searching 'COVID-19' and 'Malaysia' in the search terms in Google Scholar site. This paper also offers different insights as text modelling technique is employed to identify topics or scopes of COVID-19 research related to Malaysia that have been conducted. Through this approach, an alternative perspective is retrieved by analysing a form of unstructured data, which is text.

Data and Methods

Web Scraping Data

In the era of Industrial Revolution 4.0 and with the advent of technology, it is possible to mine data from various sources. Nowadays, data are not only in conventional form, structured and nicely tabulated; but it can be in the form of unstructured such as text and audio. Web scraping is a process to extract the data from website, largely in text form and transform this data to a structured form then stored in a database or as spreadsheet (Sirisuriya 2015).

Data used in this paper were scraped on 13th January 2021 from Google Scholar site using Octoparse. Based on Ahamad et al. (2007), Octoparse is a web crawler software that can scrape any open data from almost all websites. The search conducted in the site is using two terms which are 'COVID-19' and 'Malaysia'. The original search resulted on 681 papers in which we then filter these papers to eliminate papers without 'COVID-19' and 'Malaysia' in the title. We also filtered out papers that are not published in journals or proceedings. After the filtration process, only 134 papers that fulfilled the scope of our study were preserved. All of these papers were published in 2020. The variables extracted are title of paper, name of journal, publisher, number of citations, and affiliations.

Descriptive Analysis

Organizing and summarizing data into an interpretable form is the objective of descriptive analysis (Holcomb 1998). Describing the data is often achieved through visualization and structural form such as table. The choices of visualizations are depending on the types of data and the information that we want to convey. The data can also be summarized through statistics such as mean, median, and variance. Among examples of conventional choices of visualizations are pie chart for composition and histogram for distribution. A collection of fresher looks for visualization purposes can be found in Wickham (2011) such as waffle chart for composition and violin plot for distribution.

TOPIC MODELLING

Finding the underlying themes of text documents is the aim of topic modelling. Commonly used technique in topic modelling, the Latent Dirichlet Allocation (LDA) model is considered as probabilistic. As explained in Onan et al. (2016), this is due to the fact that each document is represented as a random mixture of latent topics and each topic is represented as a distribution over fixed set of words. In LDA, each document is considered to be constructed from a number of topics; and each topic has a mixture of words. Therefore, the generative process of document based on LDA is considered hierarchical with these layers: For each document w, choose N from Poisson (μ). N is the number of words. Choose θ from Dirichlet distribution with parameter α. Parameter θ here represents the proportion of a topic in the document. For each N words in w_n, choose the topic, z_n from a multinomial distribution with parameter θ as generated in 2. Also choose a word, w_n *from* $P(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic. A corpus is a collection of M documents, denoted by D.

Based on Zamzuri (2021), the probability density function for Poisson, Dirichlet and multinomial distributions are presented in (1) - (3).

$$f(N|\mu) = \frac{e^{-\mu}\mu^{x}}{x!} \tag{1}$$

$$d(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1} \tag{2}$$

$$g(z_1, z_2, \ldots, z_N|\theta) = \sum_{i=1}^{N}(z_i)!\prod_{i=1}^{N}\frac{\theta_i^{z_i}}{z_i!} \tag{3}$$

Since the LDA is a three layers hierarchical Bayesian model, the conditional probability density is proportional to the product of likelihood and priors as shown in (4).

$$P(\theta, z, w|\alpha, \beta) = P(\theta|\alpha)\prod_{n=1}^{N}P(z_n|\theta) + P(w_n|z_n, \beta) \tag{4}$$

The probability of a corpus is determined by taking the product of the marginal probabilities of single documents, as shown in (5).

$$P(D|\alpha, \beta) = \prod_{d=1}^{M}\int P(\theta_d|\alpha)\prod_{n=1}^{N} d\ \sum_{zd_n}P(z_{dn}|\theta_d) \tag{5}$$
$$+ P(w_{dn}|z_{dn}, \beta)d\ \theta_d$$

The number of topics, k in (2) is often determined first before the LDA model can be fitted to the data. Among approaches suggested in the literature to determine the number of topics are perplexity and coherence score (Zhou et al. 2015). The decision on the number of topics also depend on humans' judgement to maintain the semantic meaning of the terms (Cai et al. 2019). Overall, the mixture of machine automated and human are the best way to determine the number of topics.

Once the topics for each document are identified along with the related words, the importance of these words to the topic can be computed, β_{ij} represents the probability of the i-th topic containing the j-th word. The estimation of parameters in this model is performed using probability simulations or known as Gibbs sampling in Bayesian framework. Simulation is a common technique in statistical analyses as can be found in Zamzuri and Gwee (2020) and Zamzuri et al. (2018).

RESULTS AND DISCUSSION

DESCRIPTIVE STATISTICS

Number of citations

Based on number of citations, there are two papers (1.49%) with more than 100 citations. Both papers are focusing on the socio-economy impact of COVID-19 in Malaysia. The first paper with 174 citations is by McLaren et al. (2020) with the paper entitled 'COVID-19 and Women's Triple Burden: Vignettes from Sri Lanka, Malaysia, Vietnam and Australia' published in a journal named Social Sciences by Multidisciplinary Digital Publishing Institute (MDPI). As indicated clearly in the title, this paper is not solely focusing on Malaysia, but more onto the comparison between four countries listed. The issue discussed in the paper is the impact of COVID-19 to women's burden, hence it is published in a social sciences journal. The second paper with 173 citations is by Azlan et al. (2020) with title 'Public knowledge, attitudes and practices towards COVID-19: A cross-sectional study in Malaysia'. This paper is authored by a group of Malaysians researchers from Universiti Kebangsaan Malaysia (UKM), Universiti Putra Malaysia (UPM), Universiti Malaysia Perlis (UniMAP) and Sunway University Malaysia. The paper is published in a journal named PLOS which is stands for Public Library of Science. It can be perceived that both papers are focusing on impact and consequences of COVID-19 to communities localized by the country.

Another stream of interests for COVID-19 related research papers surely is in the medical and clinical analysis. This type of papers focusing on Malaysian communities is not being cited as many as the two papers previously mentioned. This is understandable as there are

plenty of papers on clinical trials and medicals published by international researchers, and often these types of papers are not focusing on the localities by countries.

Next, we look into papers with number of citations between 20 and 100 and authored by Malaysians. Ten papers (7.46%) were recorded with this characteristics, in which the details on the frequency of these papers based on the topics or field of the papers with the affiliation of the main author, categorized as public or private institutions are presented in Table 1. The numbers are relatively the same as more papers were authored by public universities and government bodies. More papers are published in science and technology field compared to social sciences. Other than that, we also observed that 64 papers published (47.76%) have not been cited yet by any other publications.

TABLE 1. The top 10 most cited papers based on institutions and fields

Field / Institution	Public University / Government body	Private University
Social Sciences	3	1
Science and Technology	4	2

Publishers and Publication Language

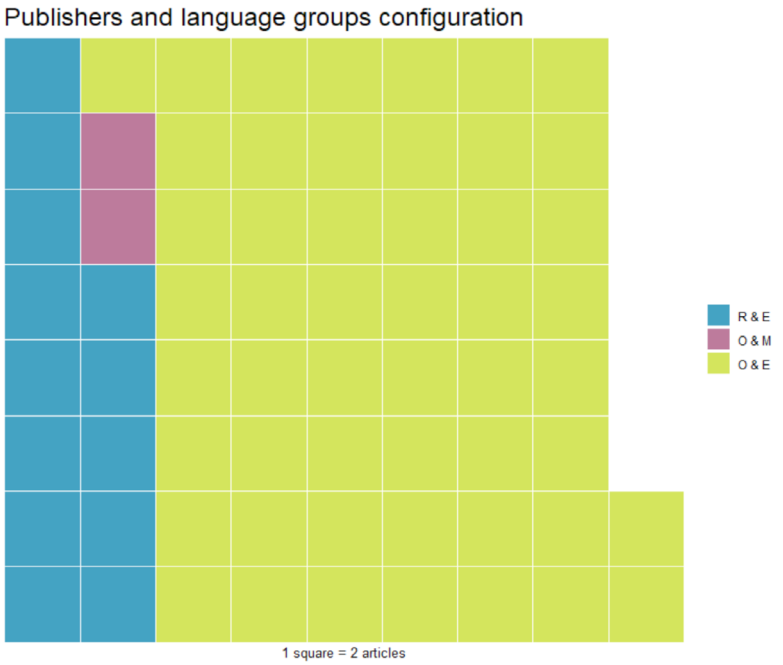

FIGURE 1. The configuration of papers based on publishers and language

As mentioned before, the scraped data also includes the information of the papers' publishers. We then categorized these publishers into two groups, reputable (R) and others (O). The list of reputable publishers used for this categorization can be found in the List of Reputable Publishers accessed from the Tun Seri Lanang Library, UKM website (2021). Among the publishers listed as reputable are Springer, Elsevier, and Taylor & Francis. We also classify the papers based on the language, either Malay (M) or English (E). Figure 1 presents the waffle chart of the papers' configuration based on publishers and language. Each square in this figure represents 2 articles and the colour represents the publishers and language group as follows, Reputable and English (R&E), Other and Malay (O&M) and Other and English (O&E). We can see that most of the papers specifically 104 papers (77.61%) were written in English and published by other publishers. There are only four publications (2.99%) written in Malay and 26 papers (19.4%) were published by reputable publishers.

TOPIC MODELLING

We perform two separate analyses of LDA of the data set. The first one is using the text data of the paper's titles followed by the text data of the journal's names. For both analyses, the optimum number of topics were identified based on the rules as explain in the methods section.

The Paper Title

To label the topics obtained from the LDA analysis, we plotted the word cloud to show the top words according to each topic, as exhibited in Figure 2. The first topic is focusing on the impact of COVID-19 to various fields, largely on the education as related words such as student, university, learn and teach appear among the top words. Other words categorized on this topic are media, tourism, and psychology. For the second topic, it is apparent that the titles of the papers are largely related to the Movement Control Order (MCO), the action taken by the Government of Malaysia in order to control the spread of the virus. Whereas for the last topic, the string of words such as response, study, infect, clinic, trend and outbreak suggests that the papers categorized into this topic is focusing on the clinical and health study. It is worth to mention that the words 'business' and 'strategy' are also categorized into the last topic.

Since there are three topics identified in Figure 2, we now choose two papers randomly for each topic. The titles of these six papers are given in Table 2. Figure 3 depicts the documents to topics probabilities for six chosen papers. Referring to Figure 3, the plot represents the document to topics probabilities. As can be seen in this figure, the *x*–axis (value) is the probabilities and the *y*-axis are the three topics as identified. The colour of the bar represents the six papers or referred as document in the plot. For the first paper, the probability of this document to belong to the third topic (respons.studi.infect.clinic. outbreak) is very high which is more than 0.75. When we refer back to Table 2, the title for the first paper clearly indicates that this paper belongs to clinical study, hence the document to topic probability is high for the third topic. Another example is paper 2 in which the probabilities for this document to belong to Topic 1 (impact.learn.industri. student.experi) and Topic 2 (control.movement.case.order. impact.) are reasonably the same, in which 0.467 for Topic 1 and 0.5 for Topic 2. Since the probability value for Topic 2 is slightly higher than Topic 1, this paper is categorized into the second topic. When we look at the title, it is comprehensible that the paper can be categorized into both topics since this paper studies on MCO and its impact on tourism and hospitality. These examples illustrate how the LDA works through identifying and comparing probabilities of word to document and document to topics computed by layers of hierarchical models under Bayesian framework.

FIGURE 2. The word clouds for three topics based on the papers' title

Table 3 provides the summary statistics of the papers focusing on the number of citations and reputable publishers. Out of the three topics, the topic with the highest number of papers is clinical and business strategy. For all three topics, the average number of citations is 6.5, 4.9 and 7.7 respectively for each topic. When we look into the median values for the number of citations, the value is zero and large variances are also observed for all three topics. Both mean and median measuring the central tendency of the data and the variance measures the dispersion or variation of data points from its mean. Based on the findings in Table 3, the median for the number of citation is 0 since 40 to 50% of the papers for all three topics have not been cited yet. The large value of variances is also due to the same reason in addition to the fact that only few papers have high citations for each topic. For example, in Topic 1, there is only one paper with 174 citations, four papers with more than 10 citations and 18 papers with zero citation. A similar pattern is also observed for Topic 3 in which with the presence of a paper with 173 citations and 29 out of 54 papers (53.7%) have not been cited yet, hence the variance is considerably large at the value of 634.489. Table 3 also show that, for all three topics, the percentage of papers published by the reputable publishers is less than 30%. Based on Tahamtan et al. (2016), the number of citations depend not only to paper and journal related factors but also authors' reputation. Since all of these papers were published just last year, in 2020; and less than 30% were published by reputable publishers, hence the fact that around 50% of the papers have not been cited yet is plausible.

TABLE 2. The title of six selected papers

Paper	Title
1	Laboratory-confirmed case of Middle East respiratory syndrome coronavirus (MERS-CoV) infection in Malaysia: preparedness and response, April 2014
2	The Movement Control Order (MCO) for covid-19 crisis and its impact on tourism and hospitality sector in Malaysia
3	Air quality status during 2020 Malaysia Movement Control Order (MCO) due to 2019 novel coronavirus (2019-nCoV) pandemic
4	COVID-19 Outbreak in Malaysia
5	Psychological impact of COVID-19 and lockdown among university students in Malaysia: Implications and policy recommendations
6	Clinical features of patients with rheumatic diseases and COVID-19 infection in Sarawak, Malaysia

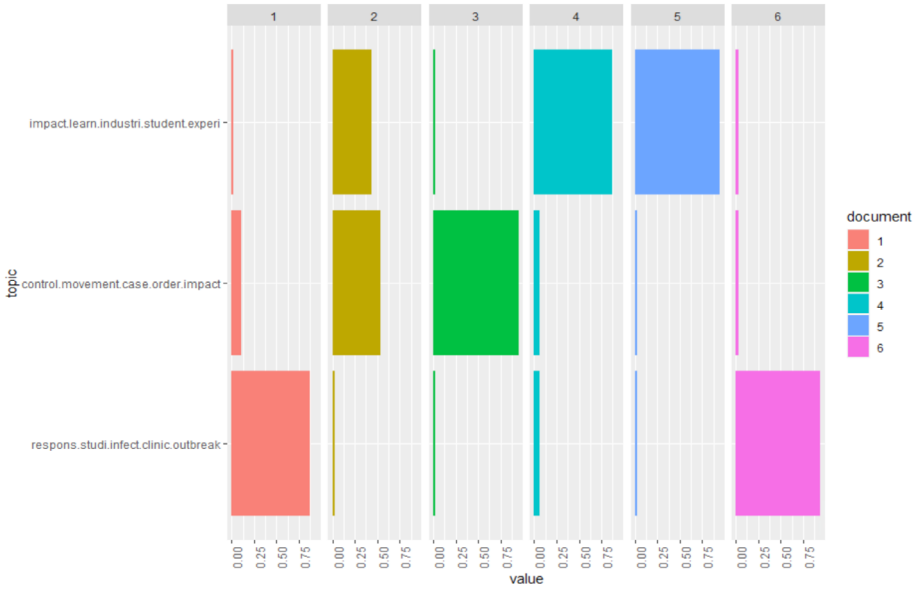

FIGURE 3. The document to topic probabilities plot for LDA based on the papers' title

TABLE 3. Summary statistics of the three topics for LDA based on the papers' title

	Topic 1 (impact)	Topic 2 (MCO)	Topic 3 (clinical)
Mean	6.537	4.878	7.655
Median	0	0	0
Variance	745.605	98.010	634.489
Maximum	174	43	173
Number of articles with 0 citation (%)	18 (45.0)	17 (42.5)	29 (53.7)
Number of articles published by reputable publishers	7 (17.5)	7 (17.5)	13 (24.1)
Total number of articles	40	40	54

The Journal Name

We perform the same analysis to the data of journal's names for these 134 papers. Three topics are identified as optimum and the top words for these topics are presented in Figure 4. The first topic is related to medical since this word along with 'disease' and 'infectious' are in the top positions. Also categorized into this topic are other words not related to medical such as 'language' and 'multidisciplinary', but with lower frequencies. The second topic's top words are 'Asean', 'European', 'business' and 'education'. Hence, this topic can be labelled as business and education. The last topic exhibits top words such as 'health', 'public', 'social' and 'science'. Medical related words are also categorized here such as 'medicine' and 'medical' but not as frequent as in the first topic. Therefore, we label the third topic as public health and social science. Based on Shelton et al. (2017), the contributions of social science to the public health aspect need to be recognized due to its significance and relevance. Hence, the third topic marries this two fields together, clinical and social science, basing on the role and impact of health study to the society.

Table 4 displays the journal names of another six selected papers, while Figure 5 depicts the document to topic probabilities plot. The allocation of the document to the topic is apparent when we observe the first paper is allocated to the first topics. The computed probabilities for the first two papers are highest for the second topic as the words of 'public health' can be observed in the journal name. Papers 3 and 6 have education related words, hence these papers are allocated to the third topic. A different pattern observed for the fifth paper in which the probabilities for the first and second topics are almost equal. It is explicable since the word 'medical' is belong to both topics. This situation illustrates the situation in which when there are two topics tied for the top position, then the algorithm will randomly assign the document to one of the top topics.

FIGURE 4. The word clouds for three topics based on the journals' name

TABLE 4. The journal names of six selected papers

Paper	Journal name
1	International Journal of Environmental Research and Public Health
2	Malaysian Journal of Public Health Medicine
3	Journal of Postgraduate Current
4	Medical Journal Malaysia
5	Malaysian Journal of Medical Sciences
6	Asian Education and Development Studies

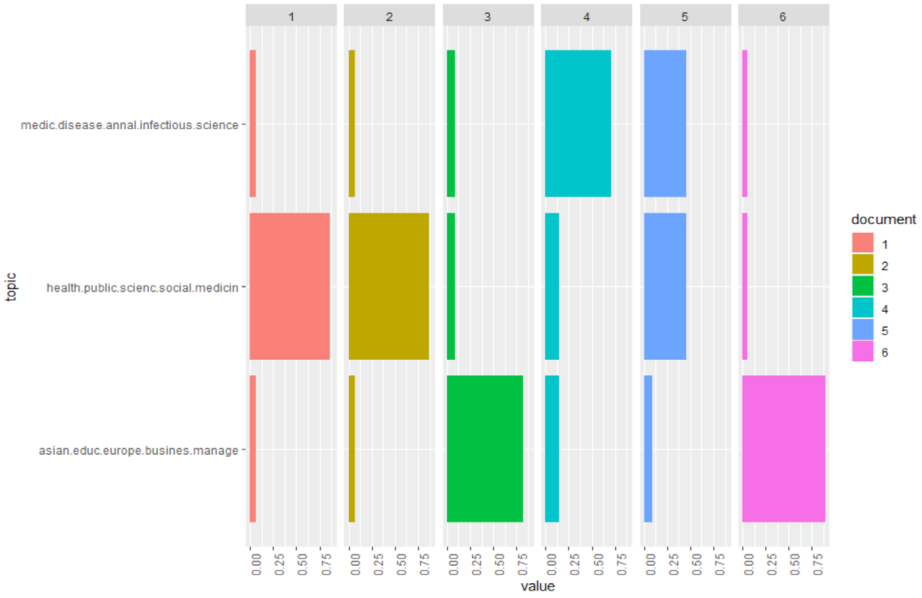

FIGURE 5. The document to topic probabilities plot for LDA based on the journals' name

Table 5 provides the summary statistics for the three topics. For the number of citations, the mean for topic 2 is the highest since there is a paper in this topic has 174 citations. Similar pattern as observed in Table 3 before, in which the median values are smaller than the mean for all three topics. Based on Pham-Gia and Hung (2001), both mean and median are the measure for the central tendency of the data, hence these measures are comparable; in which median is more robust to the presence of outliers. Since all topics consists of at least one paper with high number of

citations; 43 for Topic 1, 174 for Topic 2 and 50 for Topic 3, the results implicate that the mean value is higher than the median. We also observe that the variances for all three topics are considerably large. Looking at the percentage of papers with zero citation, papers that belong to Topic 1 have the smallest percentage in which only 37.5% of these papers have not being cited by any publications yet. The percentage for number of articles published by reputable publisher are comparatively the same, around 20% for all three topics. Topic 2 which is related to public health are the topic with the least number of articles which is 38.

TABLE 5. Summary statistics of the three topics for LDA based on the journals' name

Number of citations	Topic 1 (medic)	Topic 2 (pub. health)	Topic 3 (Asian edu.)
Mean	3.776	12.103	4.735
Median	0	1	0
Variance	62.886	1481.726	148.407
Maximum	43	174	50
Number of articles with 0 citation (%)	18 (37.5)	19 (50.0)	27 (56.3)
Number of articles published by reputable publishers	10 (20.8)	9 (23.7)	8 (16.7)
Total number of articles	48	38	48

To sum up the results of the two LDA analyses, we can see that research on COVID-19 emphasizing on Malaysia scenario follow the global pattern in which majority of the paper focusing on clinical study since COVID-19 itself is a medical issue. Due to the serious consequences of the virus, in which not only health is effected, the impacts of it are significant economically and socially. Hence, a topic focusing on the impact of COVID-19 was identified from the analysis. We also identify that there is a cluster of papers look into the MCO action taken by the government. This is a local issue as such name, the 'Movement Control Order (MCO)' is specifically used in Malaysia scenarios. Based on the findings, more papers related to MCO are expected since the second MCO have been implemented by the government on 22nd January 2021. Papers that study on the impact of COVID-19 that highlighting the economy and education issues are another popular option for the study.

For the journal title, publishing in journal with medical related names provides advantages to be cited more and has higher chances to be published by reputable publishers. Based on the second LDA results, we know that other than journals with medical terms in the name, journals related to public health and social sciences are also publishing articles related to COVID-19 in Malaysia. Another valuable input obtained from the result is journals with the word 'ASEAN' in the name, and focusing on business and education also provide huge potential to be options for the researchers to publish their findings related to the issue.

Conclusion

This paper conducted a bibliometric analysis to research conducted related to COVID-19 focusing on Malaysia cases. Taking a different approach than the conventional bibliometric analyses, Latent Dirichlet Allocation (LDA) technique was employed in order to extract semantic content and identify topics underlying under the paper's titles and the journal's names. Data used for this study was scraped from Google Scholar search engine using Octoparse. In just one year, 134 papers related to COVID-19 in Malaysia scenarios have been published.

The LDA analyses show that three topics were identified for both paper's titles and journal's names documents. Among the total of six topics, papers focusing on clinical studies are the ones that receive more attention reflected by the number of citations. A topic that represent locality is also identified which is the MCO topic. Majority of the papers tend to discuss the COVID-19 impact on various fields with the popular choices are business and education. This study contributes to describe and summarize the pattern and tendency of academic papers related to COVID-19 in Malaysia. Perhaps the findings from this paper may help researchers to identify the gap, plan and further strategize their future work on this issue. Discussing on the directions for future research, we suggest that the data should be scraped from other sources such as Scopus and Clarivate Analytics databases, in order to perform a bibliometric analysis focusing on the reputable and indexed papers. Other topic modelling techniques can also be considered and added as comparisons to the LDA technique used in this paper. Since MCO is identified as one of the prominent topic for COVID-19 research in Malaysia, analyses focusing only to papers studying effects of MCO should be done in order to identify categories of interest for the MCO action.

ACKNOWLEDGEMENTS

We would like to thank Universiti Kebangsaan Malaysia for providing financial means of this study through the grant GUP-2018-110.

REFERENCES

Ahamad, D., Mahmoud, A. & Mahmoud, M. 2017. Strategy and implementation of web mining tools. *International Journal of Innovative Research in Advanced Engineering* 12(4): 1-7. doi: 10.26562/IJIRAE.2017.DCAE10081.

Aristovnik, A., Ravšelj, D. & Umek, L. 2020. A bibliometric analysis of COVID-19 across science and social science research landscape. *Sustainability* 12(21): 9132.

Azlan, A.A., Hamzah, M.R., Sern, T.J., Ayub, S.H. & Mohamad, E. 2020. Public knowledge, attitudes and practices towards COVID-19: A cross-sectional study in Malaysia. *PLoS ONE* 15(5): e0233668.

Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A.A., Salhab, H., Fares, M. & Khachfe, H.H. 2020. A bibliometric analysis of COVID-19 research activity: A call for increased output. *Cureus* 12(3): e7357. doi: 10.7759/cureus.7357.

Cai, C.W., Linnenluecke, M.K., Marrone, M. & Singh, A.K. 2019. Machine learning and expert judgement: Analyzing emerging topics in accounting and finance research in the Asia-Pacific. *Abacus* 55(4): 709-733.

Deng, Z., Chen, J. & Wang, T. 2020. Bibliometric and visualization analysis of human coronaviruses: Prospects and implications for COVID-19 research. *Front. Cell. Infect. Microbiol.* 10: 581404. doi: 10.3389/fcimb.2020.581404.

Elengoe, A. 2020. COVID-19 outbreak in Malaysia. *Osong Public Health and Research Perspectives* 11(3): 93-100.

Fan, J., Gao, Y., Zhao, N., Dai, R., Zhang, H., Feng, X., Shi, G., Tian, J., Chen, C., Hambly, B.D. & Bao, S. 2020. Bibliometric analysis on COVID-19: A comparison of research between English and Chinese studies. *Frontiers in Public Health* 8: 477. doi: 10.3389/fpubh.2020.00477.

Holcomb, Z.C. 1998. *Fundamentals of Descriptive Statistics*. New York: Taylor & Francis.

McLaren, H.J., Wong, K.R., Nguyen, K.N. & Mahamadachchi, K.N.D. 2020. COVID-19 and women's triple burden: Vignettes from Sri Lanka, Malaysia, Vietnam and Australia. *Social Sciences* 9(5): 87.

Onan, A., Korukoglu, S. & Bulut, H. 2016. LDA-based topic modelling in text sentiment. Classification: An empirical analysis. *International Journal of Computational Linguistics and Applications* 7: 101-119.

Pham-Gia, T. & Hung, T.L. 2001. The mean and median absolute deviations. *Mathematical and Computer Modelling* 34: 921-936.

Shelton, R.C., Hatzenbuehler, M.L. & Metsch, L.R. 2017. Future perfect? The future of the social sciences in public health. *Front. Public Health* 5: 357. doi: 10.3389/fpubh.2017.00357

Sirisuriya, S.C.M. 2015. A comparative study on web scraping. *Proceedings of 8th International Research Conference of KDU*. General Sir John Kotelawala Defence University. pp. 135-140.

Tahamtan, I., Afshar, A.S. & Ahamdzadeh, K. 2016. Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics* 107: 1195-1225.

Verma, S. & Gustafsson, A. 2020. Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. *Journal of Business Research* 118: 253-261.

Wickham, H. 2011. ggplot2. *Interdisciplinary Reviews Computational Statistics* 3(2): 180-185.

Zamzuri, Z.H. 2021. Underreporting traffic accidents in Malaysia - A sentiment analysis. *International Conference on Mathematics, Statistics and Their Applications* 36: 01015.

Zamzuri, Z.H. & Gwee, J.H. 2020. Comparing and forecasting using stochastic mortality models: A Monte Carlo simulation. *Sains Malaysiana* 49(8): 2013-2022.

Zamzuri, Z.H., Sapuan, M.S. & Ibrahim, K. 2018. *The extra zeros in traffic accident data: A study on the mixture of discrete distributions. Sains Malaysiana* 47(8): 1931-1940.

Zhaou, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y. & Zou, W. 2015. A heuristic approach to determine an appropriate number of topics in topic modelling. *BMC Bioinformatics* 16: S8. https://doi.org/10.1186/1471-2105-16-S13-S8.

Department of Mathematical Sciences
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: zamira@ukm.edu.my

Received: 29 January 2021
Accepted: 28 April 2021