International Journal of Multidisciplinary Research and Development
www.allsubjectjournal.com
Online ISSN: 2349-4182, Print ISSN: 2349-5979
Received: 05-05-2022, Accepted: 21-06-2022, Published: 06-07-2022
Volume 9, Issue 7, 2022, Page No. 43-49

Text analytics on Covid-19 sentiment in Malaysia using K-means clustering approach

Ashraff Ruslan[1*], Norshahida Shaadan[1], Muhammad Khalis Abdul Karim[2]
[1] Center for Statistics and Decision Science, Faculty of Computer and Mathematical Sciences, University Teknologi MARA, Shah Alam, Selangor, Malaysia
[2] Department of Physics, Faculty of Science, Universiti Putra Malaysia, Selangor, Malaysia

Abstract
Text clustering has been acknowledged as a useful technique in clustering the unstructured data. One of the main resources of unstructured data is digital media where people usually obtained from the internet. For example, the English Wikipedia includes 6,295,065 articles and has averages of 593 new articles per day. Hence, this study concerns on the web mining technique for clustering text data related to COVID-19 (Cov-19) sentiment from digital media. 20 articles related to Malaysian sentiment on Cov-19 were extracted using phyton language. The feature selection and assigning weightage to the terms were performed using Term Frequency-Inverse Document Frequency (TF-IDF) approach. Furthermore, the k-means clustering was used for text clustering using *sklearn* module in python. This approach capable to cluster 20 articles into 6 main themes based on period where Malaysian government acted on Cov-19 issues, Cov-19 death cases, timeline of pandemic, the origin of the virus, diagnosis and treatment measure of Cov-19 and medical institution which handle the cases. In conclusion, the proposed clustering technique managed to indicate the focus content on which Cov-19 were discussed in the digital media. Hence, this approach can be further extend to more depth research area such as sentiment analysis and building corpus on other fields.

Keywords: text mining, text clustering, k-means, elbow method, cov-19 sentiment

Introduction
The advancement of digital world allows internet user to acquired latest information from electronic format, and most of this information are available as a text in news stories, technical papers, books, digital libraries, email messages, blogs, and web pages [1]. One of the most essential issues nowadays is knowledge mining, such as pattern detection or grouping of similar terms on the unstructured data, especially text. Most informatic professionals are familiar with structured data and dealing with this type of data due to its reliability and feasibility [2].

Recently, the COVID-19 (Cov-19) pandemic in Malaysia is part of a worldwide pandemic caused by coronavirus 2 (severe acute respiratory syndrome) (SARS-CoV-2). The case of Cov-19 became significantly increase right after the first imported case from Wuhan, China on January 23, 2020, reported by Singapore authority. Following from the first case, another eight close contacts were traced back and already reside in Johor, Malaysia [3]. With less than 48 h of the first case reported in Singapore, Malaysia reported its first Cov-19-positive case on January 25, 2020, due to import case from neighboring region. A few days later, the first Malaysian testing positive for Cov-19 was reported on February 3, 2020; where the person had a history of travel to a neighboring country for a business meeting, which was also attended by a delegation from China. With stringent policy at the entrance, the first wave was successfully handled by February 27, 2020, with all 22 previously reported cases being discharged from hospital. But later, other waves strike for another 8 months, and authority decide to add more restriction especially on the travelling policy up to until now. Ministry of Health (MOH) of Malaysia has overseen the mitigations and readiness to the outbreak in Malaysia. With over 2,200,000 verified Cov-19 cases, over 41,221 active cases, and over 31,487 deaths as of December 31, 2021, the country is now ranked third in Southeast Asia in terms of Cov-19 cases and deaths, behind Indonesia and the Philippines [4, 5].

In this paper, the data on the recent pandemic Cov-19 in Malaysia has been fetched from available digital media (webpage, news, Wikipedia), analyzed, and clustered using the unsupervised method. We have utilized Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme for the purpose. Basically, IDF method will assign weight to each word according to the importance of word instead of determining the frequency of the word in the documents Since the frequency of words tend to incorrectly emphasize words only based on the more frequent word exist, IDF diminish the weight of the word that occurs very frequent and increases the weight of the term that occur rarely.

Several researchers had performed the text clustering using K-means clustering as it will actualize the dataset [6, 8]. K-means algorithm is an unaided calculation that takes various information focuses and bunch them into a k number of groups. Here, k indicates the number of bunches, the number of groups that will be expanded into three bunches. In k-means calculation, information focuses a plotted over a disperse chart, and k number of bunches are set. A calculation will be done to the quantity of cycle set to discover the information directs closest toward the centroid's dependent on Euclidean separation. After the most extreme emphasis, the bunch of information focuses on the centroids will be the last groups. K-means clustering is a very well-known clustering technique and algorithm [9, 10]. The grouping/mastery of computer and internet usage indicators can be made grouping to facilitate analysis. Text data in Malaysian context is still growing and there are so many areas and platform of information's sources can be explored [11]. Examples of focus area when exploring text data are text analysis, sentiment analysis from social media platform and web pages and topic detection from text document stored in hardcopy or softcopy. However, the main drawback from the technique includes on challenges in acquiring the right data, reliable tools, and suitable algorithm for the exploration. Hence, this study aims to mine and clustering the text data related to Cov-19 sentiment among Malaysian perspective based on the digital media using the eligible technique such as K-means clustering.

Materials and Methods

1. Data Preparation

This study scrutinizes 20 topics of articles from Wikipedia (website address, origin country) page by using Python (website, origin country) software to automate the extraction using library. The keyword use includes 'Covid-19', 'Malaysia Covid-19' and 'Ministry of Health Malaysia'. Subsequently the text clustering was organized into four steps: (1) Data Extraction from Wikipedia page; (2) Feature Selection and Exploration; (3) Text Clustering; (4) Result evaluation. Figure 1 shows the research framework for this study.

2. Data Extraction

In this study the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme were utilized for text mining. The logarithmic scale factor for the importance level of the word also called IDF is defined as follows:

$$idf(j) = \log\left(\frac{N}{df(j)}\right) \tag{1}$$

with the weight of full weighting for j^{th} term is:

$$tf - idf(j) = tf(j) \times idf(j) \tag{2}$$

where *tf(j)* is the frequency of the j^{th} term and *idf(j)* is the logarithmic scale factor for the importance level of the word. We have extended the utilization of TF-IDF using sklearn module in Python. The exploration of the number of distinct terms contained in the data set, sklearn.feature_extraction.text. CountVectorizer were used.
The data then were vectorized using *sklearn.feature_extraction.text.TfidfVectorizer* and TF-IDF values is computed for further clustering Module of stop_words was used for stopping words removal. Using stop words is part of the cleaning and feature selection as the articles may contained words or terms that are not meaningful for the clustering. This cleaning is necessary to retained only meaningful words/term according to the stop words library. For this study, the build-in stop words library is used in the process of elimination of the meaningless words and terms.

3. Text Clustering

The K-means algorithm is applied on vector representation of selected articles that was obtained in the earlier section using sklearn.cluster.Kmeans. Here k de-notes as the number clusters ~ initialized the K number of centroids in the data with range between 2 to 10 with the maximum iteration is set to 200. The distance of each word were calculated using Squared Euclidean Distance. Figure 1 shows sample of source code in computing the K-means clustering in this study.

```
from sklearn.cluster import KMeans
Sum_of_squared_distances = []
K = range(2,10)
for k in K:
    km = KMeans(n_clusters=k, max_iter=200, n_init=10)
    km = km.fit(X)
    Sum_of_squared_distances.append(km.inertia_)
```

Fig 1: Sample code for K-means clustering

The number of clusters in the data set were decided using Elbow method. The chart of sum of squared distances per number of clusters were plot and the elbow of the curve was observed to decide on the optimal number of clusters. The sample code for performing Elbow Method is shown in Figure 2.

```
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

Fig 2: Sample of source code in plotting Elbow Method

The optimal number of clusters obtained from the above plot is then used for final clustering and converted into data frame to view the results. For better rep-presentation, the result of clustering is transformed into word cloud by using the codes as shown in Figure 4. Furthermore, the performance of the clusters was finally measured using silhouette score from *sklearn.metrics.silhouette_score* and *sklearn.metrics.calinski_harabasz_score* [12, 13].

```
from wordcloud import WordCloud
result={'cluster':labels,'wiki':wiki_lst}
result=pd.DataFrame(result)
for k in range(0,true_k):
   s=result[result.cluster==k]
   text=s['wiki'].str.cat(sep=' ')
   text=text.lower()
   text=' '.join([word for word in text.split()])
   wordcloud = WordCloud(max_font_size=50, max_words=100, background_color="white").generate(text)
   print('Cluster: {}'.format(k))
   print('Titles')
   titles=wiki_cl[wiki_cl.cluster==k]['title']
   print(titles.to_string(index=False))
   plt.figure()
   plt.imshow(wordcloud, interpolation="bilinear")
   plt.axis("off")
   plt.show()
```

Fig 3: Sample of source code in plotting Word Cloud

Results and Discussion

In this work, 20 articles were successfully extracted from digital media such as Wikipedia and the text from each article is stored in the *wiki_lst* in Python and become the object of the study. From the CountVectorizer, the result of the extraction is transformed into a vector of 20 rows of articles and 8,824 distinct terms. The English stop words such as “is”, “a”, “the”, “such” (all 318 English stop words) then applied to the vector and resulted to 8,569 terms left in the vector. Parts of the term frequencies contained in articles fetched are tabulated in Table 1 and Table 2.

As indicate from Table 2, the Count Vectorizer provide number of frequencies with respect to index of vocabulary. The term “000” had appeared 36 times in the “COVID-19 pandemic in Malaysia” and 25 times in “Timeline of the COVID-19 pandemic in Malaysia (…” for instance. As for the last term “**เยวชนปลดแอก**” had appeared one time in the “Protests over responses to the COVID-19 pandemic” article and non for other articles. The term “000” for example, provides no meaning for the reader and need to identify which terms that are meaningful despite having frequently appeared in the articles. This is where the TF-IDF plays a vital role in assigning weightage of each term in the article. Thus, the term is represented in a numeric vector and ranked from the highest score to the lowest one. The score is tabulated with the top 10 terms resulted from the TF-IDF calculation as presented in Table 3. The terms that appear practically everywhere in one article and is not found in any other articles has a high degree of uniqueness on that article, and thus a high TF-IDF weight was observed in Table 3.

Table 1: Count Vectorizer of terms in articles (including stop words)

No.	Articles/Terms	00 (1st)	000 (2nd)	…	état (8,823rd)	เยาชนปลดแอก (8,824th)
1.	COVID-19 pandemic in Malaysia	1	36	...	0	0
2.	Timeline of the COVID-19 pandemic in Malaysia (...	1	25	...	0	0
3.	COVID-19 pandemic in Sabah	0	18	...	0	0
4.	Protests over responses to the COVID-19 pandemic	0	18	...	1	1
5.	COVID-19 vaccination in Malaysia	0	15	...	0	0
6.	COVID-19 pandemic in Asia	1	14	...	1	0
7.	COVID-19 pandemic in Singapore	0	12	...	0	0
8.	Timeline of the COVID-19 pandemic in Malaysia (...	0	7	...	0	0
9.	Social impact of the COVID-19 pandemic in Malaysia	0	5	...	1	0
10.	2020 Tablighi Jamaat COVID-19 hotspot in Malaysia	0	4	...	0	0
11.	COVID-19	0	4	...	0	0
12.	COVID-19 pandemic in Selangor	0	3	...	0	0
13.	COVID-19 pandemic in Sarawak	0	3	...	0	0
14.	COVID-19 pandemic in Brunei	0	2	...	0	0
15.	Investigations into the origin of COVID-19	0	1	...	0	0
16.	Impact of the COVID-19 pandemic on politics in	0	0	...	0	0
17.	Timeline of the COVID-19 pandemic in Malaysia	0	0	...	0	0
18.	List of deaths due to COVID-19	0	0	...	0	0
19.	Timeline of the COVID-19 pandemic	0	0	...	0	0
20	COVID-19 pandemic in Jharkhand	0	0	...	0	0

Table 2: Count Vectorizer of terms in articles (excluding stop words)

No.	Articles/Terms	00 (1st)	000 (2nd)	…	état (8,568rd)	เยาชนปลดแอก (8,569th)
1.	COVID-19 pandemic in Malaysia	1	36	...	0	0
2.	Timeline of the COVID-19 pandemic in Malaysia (...	1	25	...	0	0
3.	COVID-19 pandemic in Sabah	0	18	...	0	0
4.	Protests over responses to the COVID-19 pandemic	0	18	...	1	1
5.	COVID-19 vaccination in Malaysia	0	15	...	0	0
6.	COVID-19 pandemic in Asia	1	14	...	1	0
7.	COVID-19 pandemic in Singapore	0	12	...	0	0
8.	Timeline of the COVID-19 pandemic in Malaysia (...	0	7	...	0	0
9.	Social impact of the COVID-19 pandemic in Malaysia	0	5	...	1	0
10.	2020 Tablighi Jamaat COVID-19 hotspot in Malaysia	0	4	...	0	0
11.	COVID-19	0	4	...	0	0
12.	COVID-19 pandemic in Selangor	0	3	...	0	0
13.	COVID-19 pandemic in Sarawak	0	3	...	0	0
14.	COVID-19 pandemic in Brunei	0	2	...	0	0
15.	Investigations into the origin of COVID-19	0	1	...	0	0
16.	Impact of the COVID-19 pandemic on politics in	0	0	...	0	0
17.	Timeline of the COVID-19 pandemic in Malaysia	0	0	...	0	0
18.	List of deaths due to COVID-19	0	0	...	0	0
19.	Timeline of the COVID-19 pandemic	0	0	...	0	0
20	COVID-19 pandemic in Jharkhand	0	0	...	0	0

As tabulated in Table 3, we can observe that the highest TF-IDF Score goes to the word ‘cases’ and followed by ‘Malaysia’ in the second rank and ‘government’ in the third places. This is expected as the articles fetched about Malaysian Covid 19 cases, and those terms are meaningful for the reader. After that, the text clustering has been performed to optimize the dataset. As observed from the Figure 5, the plot is quite linear for all 20 mined articles. However, when examine closely, a clear dent is observed at k=4 and k=6. Based on the elbow Method representation, we have decided to cluster it into 6 groups.

Table 4 and Figure 5 indicate the result of the clustering based on cluster and word cloud, respectively. Regardless of only small dent observed from Elbow Method plot, the plot allow for optimizing the number of clusters in this study. It can be observed from Table 4 that the articles were clustered into Cluster 0 up to Cluster 5. Cluster 0 included in the result as the range number of cluster is set from 0 and it has 13 number of articles falls under this cluster. Altogether there were.

Table 3: Top 10 TF-IDF score from the related articles

No.	Terms	TF-IDF Score
1.	cases	0.214275
2.	Malaysia	0.178871
3.	government	0.164912
4.	total	0.151803
5.	19	0.149690
6.	covid	0.143622
7.	covid 19	0.141599
8.	2020	0.135531
9.	minister	0.132470
10.	number	0.123500

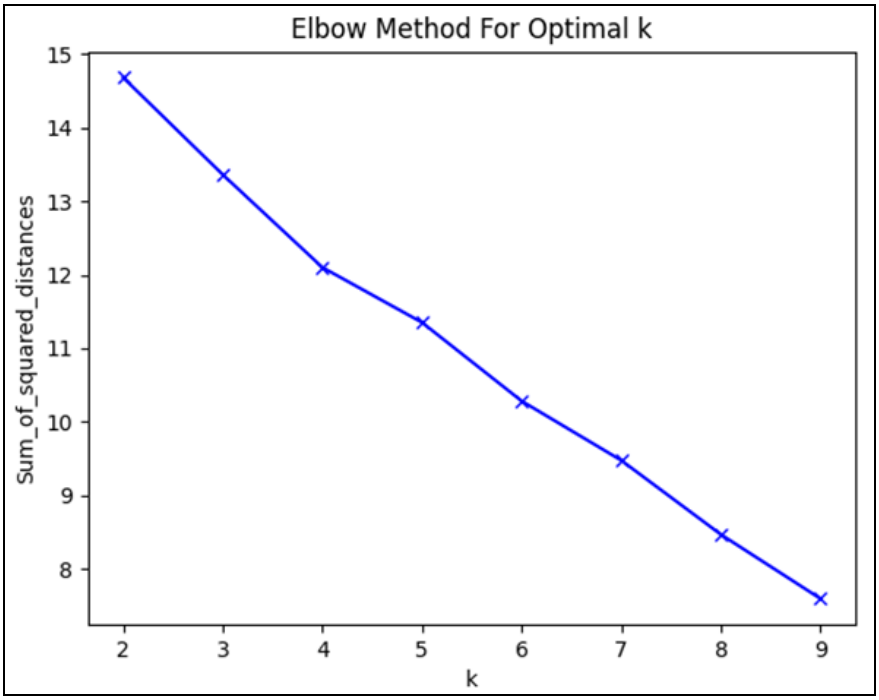

Fig 4: Sum of squared distances per number of clusters

Table 4: 6 clusters obtained in line with optimal number of clusters chosen earlier

No.	Terms	Cluster
1.	COVID-19 pandemic in Malaysia	0
2.	Timeline of the COVID-19 pandemic in Malaysia ...	0
3.	COVID-19 pandemic in Brunei	0
4.	Protests over responses to the COVID-19 pandemic	0
5.	COVID-19 pandemic in Sabah	0
6.	COVID-19 pandemic in Asia	0
7.	COVID-19 pandemic in Selangor	0
8.	Timeline of the COVID-19 pandemic in Malaysia ...	0
9.	Impact of the COVID-19 pandemic on politics in...	0
10.	COVID-19 pandemic in Singapore	0
11.	Social impact of the COVID-19 pandemic in Mala...	0
12.	COVID-19 vaccination in Malaysia	0
13.	2020 Tablighi Jamaat COVID-19 hotspot in Malaysia	0
14.	List of deaths due to COVID-19	1
15.	Timeline of the COVID-19 pandemic in Malaysia	2
16.	Timeline of the COVID-19 pandemic	2
17.	COVID-19	3
18.	Investigations into the origin of COVID-19	3
19	COVID-19 pandemic in Jharkhand	4
20.	COVID-19 pandemic in Sarawak	5

From the word cloud representation, we can see that there is no overlapping word from one cluster to another. In addition, this approach is able to cluster 20 articles into 6 main themes represent a period when Malaysian government acted on Cov-19 issues (Cluster 1), Cov-19 death cases (Cluster 2), timeline of Cov-19 (Cluster 3), the origin of Cov-19 (Cluster 4), testing and treatment measure of Cov-19 (Cluster 5) and medical institution which handle Cov-19 cases which essentially provides readers with the summary of the articles in Wikipedia that was fetched earlier.

The *silhouette_score and calinski_harabasz_score* were utilized to measure the performane index of the clustering [14, 15]. Table 6 indicate the index of the performance of the using *silhouette_score and calinski_harabasz_score*. The results from *Silhouette* score were nearly 0 indicate that the data is correctly clustered but shows high indication of overlapping cluster. Meanwhile, the results of *Calinski Harabasz* indicate that the ratio within cluster dispersion and the between cluster dispersion is fair performance.

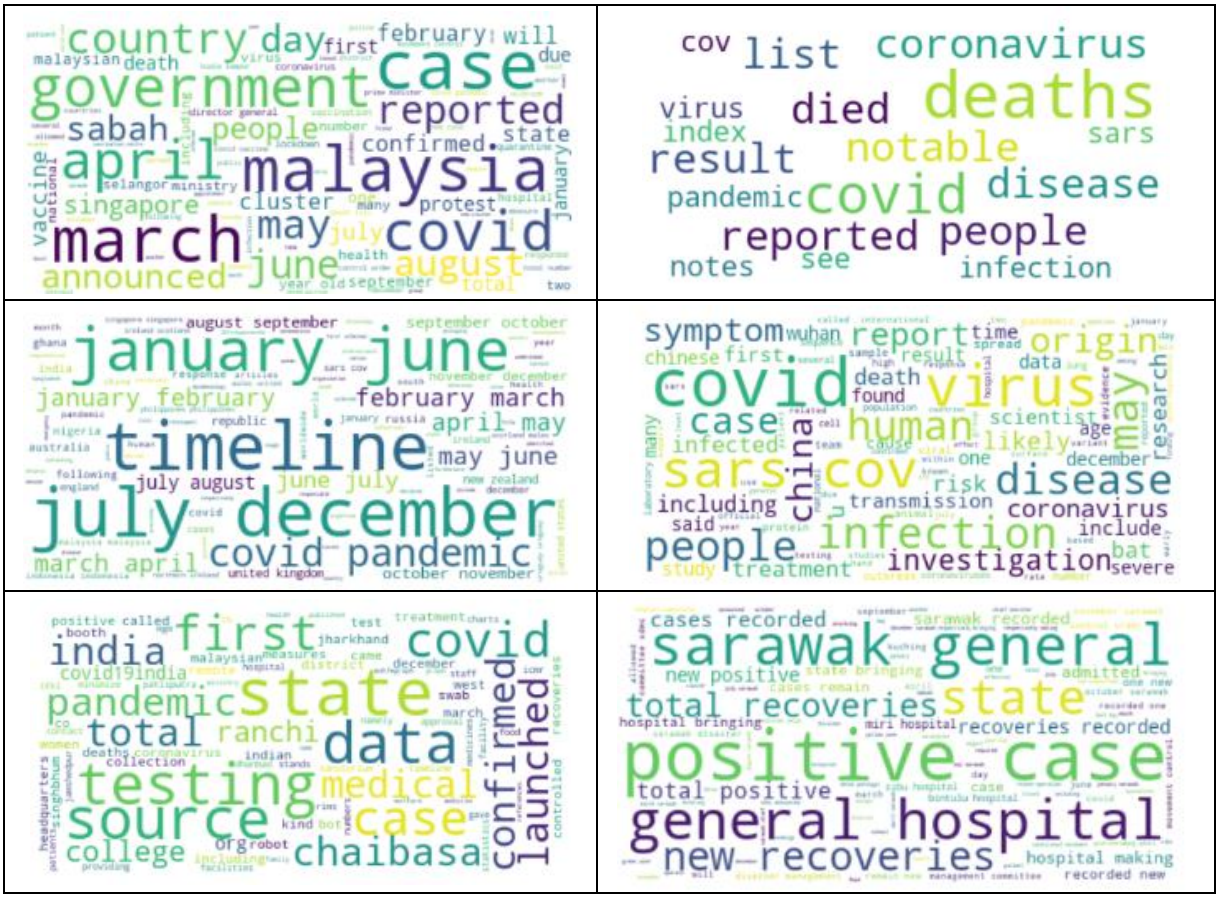

Fig 5: Word cloud of clusters related to Cov-19 in Malaysian country.

Table 6: Clusters performance obtained in line with optimal number of clusters chosen earlier.

No	Performance measurement	Results
1.	*Silhouette score*	0.0404
2.	*Calinski Harabasz score*	1.5379

There are few recommendations that can be carried out for the future study. First, this study uses a stop words technique to clean the data, hence, it is advisable to include word stemming on top of the stop words. Word stemming will help to further trim down the word into its root word to help increase the distinct between each word and remove the similarity of words from the articles. As to improve the selection of the optimal number of clusters, the iteration can be carried out with different number of K from the Elbow Method plot to obtain the improvement of the optimal cluster selection. Finally, the extraction of article can be extended and directly specify the title of the articles required of the study. This should be in line with the focus of the application of analysis of unstructured data in future to improve the performance of the cluster and reduce the cluster overlapping.

Conclusions

In this study, the articles related to Cov-19 in Malaysian country from the news in digital media were extracted using the keywords. The word cluster has able to visualize the articles contained in each cluster. The cluster performance is fair with measurement using *Silhouette score* and *Calinski Harabasz* score. As conclusion, this

approach able to be developed and can be further extend in other depth research area such as sentiment analysis and building corpus.

Acknowledgements
We would like to express gratitude to staff members of Department of Statics and Decision Science, Faculty of Computer and Mathematical Sciences, UiTM in providing input and preparation of this analysis.

References

1. Lee J, Yoon W, Kim SS, Kim D, Kim SS, So CH *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining ed J Wren Bioinformatics,2020:36:1234-40.
2. Shi H, Liu Y. Naïve Bayes vs. support vector machine: Resilience to missing data *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 7003 LNAI (Springer, Berlin, Heidelberg), 2011, 680-7.
3. Shah AUM, Safri SNA, Thevadas R, Noordin NK, Rahman AA, Sekawi Z *et al*. COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government Int. J. Infect. Dis,2020:97:108-16.
4. Foo LP, Chin MY, Tan KL, Phuah KT. The impact of COVID-19 on tourism industry in Malaysia Curr. Issues Tour,2021:24:2735-9.
5. Azlan AA, Hamzah MR, Sern TJ, Ayub SH, Mohamad E. Public knowledge, attitudes and practices towards COVID-19: A cross-sectional study in Malaysia *PLoS One,* 2020, 15.
6. Sangaiah AK, Fakhry AE, Abdel-Basset M, El-henawy I. Arabic text clustering using improved clustering algorithms with dimensionality reduction *Cluster Comput*,2019:22:4535-49.
7. Pandey A, Bhimrao B, Pandey A, Malviya AK. Enhancing test case reduction by k-means algorithm and elbow method *researchgate.net*, 2018.
8. Nainggolan R, Perangin-Angin R, Simarmata E, Tarigan AF. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method Journal of Physics: Conference Series, 2019, 1361.
9. Zhao S, Li W, Cao J. A User-Adaptive Algorithm for Activity Recognition Based on K-Means Clustering, Local Outlier Factor, and Multivariate Gaussian Distribution *Sensors*,2018:18:1850.
10. Javed Mehedi Shamrat FM, Tasnim Z, Mahmud I, Jahan N, Nobel NI. Application of k-means clustering algorithm to determine the density of demand of different kinds of jobs *Int. J. Sci. Technol. Res*,2020:9:2550-7.
11. Pejic-Bach M, Bertoncel T, Meško M, Krstić Ž. Text mining of industry 4.0 job advertisements *Int. J. Inf. Manage*,2020:50:416-31.
12. Shahapure K, International CN. I 7th and 2020 undefined 2020 Cluster Quality Analysis Using Silhouette Score *ieeexplore.ieee.org,* 2020.
13. Shuai Y, Jiang C, Su X, Yuan C, Huang X. A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy *2*020 IEEE 6th International Conference on Control Science and Systems Engineering, ICCSSE 2020, 68-71.
14. D'Silva J, Sharma U. Unsupervised Automatic Text Summarization of Konkani Texts using K-means with Elbow Method Int. J. Eng. Res. Technol,2020:13:2380-4.
15. Laxmi Lydia E, Krishna Kumar P, Shankar K, Lakshmanaprabu SK, Vidhyavathi RM, Maseleno A. Charismatic Document Clustering Through Novel K-Means Non-negative Matrix Factorization (KNMF) Algorithm Using Key Phrase Extraction Int. J. Parallel Program,2020:48:496-514.
16. Prof Mhareb, Imam Abdulrahman Bin Faisal University,Saudi, mhsabumhareb@iau.edu.sa
17. Dr Yahaya Musa, Ahmadu Bello University, Nigeria, yahaya_ms@yahoo.com
18. Dr Hanif Haspi, Ministry of Health Malaysia, hanifhaspi@gmail.com
19. Prof Lukmanda, Universitas indonesia, lukmanda.evan@sci.ui.ac.id
20. Dr Mardhiyati, UNISEL, Malaysia, mardhiyati@gmail.com