

Analysing corporate social responsibility reports using document clustering and topic modeling techniques

Nik Siti Madihah Nik Mangsor^{1,4}, Syerina Azlin Md Nasir², Wan Fairos Wan Yaacob¹, Zurina Ismail¹, Shuzlina Abdul Rahman³

¹Department of Computer Science, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Kota Bharu, Malaysia

²Department of Information Technology, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Machang, Malaysia

³Department of Information Systems, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia

⁴Maab Barakah Resources in Machang, Kelantan, Malaysia

Article Info

Article history:

Received Nov 27, 2021

Revised Mar 23, 2022

Accepted Apr 1, 2022

Keywords:

Annual report

Document clustering

Philanthropic corporate social responsibility

Textual analysis

Topic modeling

ABSTRACT

Corporate social responsibility (CSR) has become an imperative tool to address challenges and achieve sustainable growth. Realizing its impact to the society, companies are demanded to participate in sustainable development of which poverty eradication is one of it. The CSR practice, to date, is not strategically planned and executed especially when it comes into philanthropic corporate social responsibility (PCSR). This could be due to failure to identify categories of PCSR activities, limiting its effectiveness to achieve the intended outcomes. Thus, document clustering is proposed to be used to automate the pattern identification process. This study has extended document clustering by integrating the traditional document clustering application with topic modeling approach. This integrated approach enables the identification of the PCSR pattern. The analysis involved a three-year data from the annual report of the 25 CSR-award winning companies in Malaysia which involved several steps. Findings from this study revealed seven clusters that represented seven types of PCSR activities performed by the CSR-award winning companies in Malaysia. The findings offer an insight to be considered by companies in strategizing the CSR activities, particularly philanthropic responsibility in ensuring optimum impact to innovatively support the society and contribute towards poverty mitigation.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Syerina Azlin Md Nasir

Department of Information Technology, Faculty of Computer and Mathematical Sciences

Universiti Teknologi MARA Cawangan Kelantan

Machang Campus, 18500, Machang, Kelantan, Malaysia

Email: syerina@uitm.edu.my

1. INTRODUCTION

Corporate social responsibility (CSR) has emerged as an essential tool in today's dynamic business arena for addressing environmental issues and achieving long-term company growth. Several academics have proposed that CSR implemented by companies may increase sustainability as an emerging business strategy in complex competitive working environments and long-term growth and stability [1]-[4]. Since the rise of numerous CSR concerns, which have changed the nature, social, economic, and political aspects of organisations, CSR has emerged as a critical dimension for these entities to consider. According to Carroll [5], "Corporate social responsibility" is defined as "organisations' economic, legal, ethical, and philanthropic

that society has at a given time” [6], [7]. Philanthropic corporate social responsibility (PCSR) is widely accepted as the core of CSR due to its strong connection to cause-related marketing, which plays a critical role in influencing consumer's attitude toward the brand [8], [9]. As a result, many recent studies have focused on PCSR [10]. PCSR is defined as "Corporate actions that contribute to society's expectation of a good corporate citizen" [11]. PCSR activities are essential to raise the company's profile and improve its overall results [12].

Moreover, United Nations (UN) development organisations and countries are continuously requesting that companies contribute more to sustainable development goals, such as poverty eradication. In this regard, the 2030 United Nations Sustainable Development Plan stresses companies' commitment to tackle the issues of sustainable development, one of which is eradicating poverty and promoting fruitful public-private collaborations to achieve this goal [13], [14]. However, there is lacking attention from CSR efforts and sustainable management practises on societal issue of poverty reduction [15]. According to Musili [16], among the factors that influence poverty in Africa include societal and political greed, poor governance structures, spatial inequality during distribution of resources, unemployment, poor infrastructure, and political instability. To help reduce poverty, it requires strong institutions, non-marginalization of communities during resource allocation and transparency in government systems. However, it is not something that is easy to implement. The disparities can be seen in the distribution of resources by the government, agencies, individuals, and NGOs when a country is dealing with disaster. For example, Malaysia in early 2022, three states (Selangor, Pahang, and Melaka) were hit by extraordinary floods, destroyed properties and some even lost lives. This scenario attracted the attention of many parties to help provide donations. However, although various parties have come forward to provide help and donations, the distribution of donation is seen as inefficient and more focused on certain types of support. There was a dumping of donations of clothes and food in some *Pusat Pindahan Sementara* (PPS) [17], [18]. Besides, according to the media there is help to the needy that does not reach as expected due to authorities' failure to identify and oversee programme execution [19]. This problem is not a recent one, but it has been for a long time.

Furthermore, CSR-related activities are often ad-hoc in nature, resulting in the inability to foresee what further actions or activities to perform in future. The allocation of the budgets to the needy is seen as untargeted, resulting in those who are not eligible to receive it. These occurs due to unstructured management and difficulties in categorizing CSR activities that must be carried out [20]. The critical information is frequently disguised by redundancy and noise and clustering the data according to comparable characteristics is one method for quickly summarising the data for further study [21]. Therefore, in this scenario, identifying categories and types of philanthropic activities are needed to effectively manage CSR activities. Activities identification will enable the corporation to properly plan its resources for the betterment of the society, also could provide companies or other bodies with a context and reference point in the analysis, making plans, and evaluation of PCSR activities. Hence, the pattern identification process can be automated through document clustering technique. Document clustering is utilised in a wide variety of domains, including information retrieval, social media analysis, neuroscience, image processing, text analysis, and bioinformatics [21].

To reduce poverty in Kenyan rural areas, Musili [16] employed clustering algorithms to categorise rural farming households as poor or non-poor. Rahman *et al.* [22] presents a B40 clustering-based k-means with cosine similarity architecture to discover the appropriate indicators and dimensions for data-driven multidimensional poverty index (MPI) measurement. The clustering algorithm identified eight groups within the B40 group, which may aid the government in appropriately identifying members of the B40 group who are financially burdened and may have been misclassified before. Using the multinomial Naive Bayes technique from text mining, Das *et al.* [23] examined CSR reports from automobile industries. The study was able to identify the areas in which a CSR report should focus. According to Chae and Park [24], “structural topic modeling (STM)”, a machine learning-based content analysis tool, is used to understand themes or subjects from CSR-related Twitter interactions. Taran and Mirkin [25] apply the popular approach of k-means clustering on the conventional Morgan Stanley Capital International (MSCI) database and focus on four primary dimensions of CSR: environment, social/stakeholder, labor, and governance. Castellanos *et al.* [26] utilizes text data mining (TDM) to analyse the contents of Corporate Social Responsibility (CSR) reports. This analysis considers 5 CSR dimensions that are based on the sustainability accounting standards board (SASB) (2014).

Mishra *et al.* [27] applied k-means algorithm for clustering to identify the segments of the datasets. Besides, work by Sherkat *et al.* [28] proposed K-means algorithm to implement as a clustering method to identify the field of jobs people are most interested in and arrange for training sessions [29]. In addition, [30] used latent dirichlet allocation (LDA) approach for discovering topics in a textual corpus based on a semantic analysis of ‘big data software engineering (BDSE)’ job advertisements. Sutherland and Kiatkawsin [31] used LDA to extract latent topics from the corpus of Chinese short text. Thus, they were able to generate topic-

focused and less noisy features. For topic modeling and classification, [32] proposed the LDA and possibilistic fuzzy c-means (PFCM) methods. The proposed approach increased accuracy in twitter sentiment analysis by up to 3.5%. The study was done by [33] uses LDA and k-means clustering algorithm on Arabic text documents and the results reveal that the combined method results substantially better than the simple algorithm k-means. However, there is a limited number of the application of hybridization between topic modelling and document clustering being used in other domains, but not in CSR.

Therefore, this research attempts to enhance the conventional technique by proposing an integration of document clustering and topic modeling approach in categorizing CSR activities. Moreover, the purpose of this study does not only identify the hidden pattern of CSR activities which will be used as a strategic approach in planning, but also the finding could assist the government to strategize and increase participation of companies and NGOs in CSR activities for poverty mitigation programs. In this paper, an explanation of document clustering and topic modeling is given. The detailed description of both techniques taken in this study is further elaborated. It then follows the results of the analysis and finally the discussion on the implications of future research.

2. METHOD

The proposed method consists of a few steps such as, dataset collection, text pre-processing, document clustering and topic modelling. As for the implementation of this experiment, RapidMiner has been used in this simulation. This tool is prominent as data mining software. The brief description about the proposed method is presented in these subsections.

2.1. Dataset collection

This experiment used a dataset gathered from the 3-year annual reports by the 25 CSR-award winning companies in Malaysia 2020, available on the data stream [34]. An annual report is a detailed description of the activities of a corporation during the previous year. Firms released the CSR reports in order to share their initiatives and outcomes on social responsibility. The dataset of the annual report from the year 2017 to 2019 from each company was collected. Then, extracting unstructured data of the selected textual data based on CSR activities of each annual report was carried out and converted into a structured format. The documents were then collated and summarized for the cleaning process. The list of 25 CSR-Award winning companies is shown in Table 1.

Table 1. List of CSR-award winning companies

CSR-Award Winning Companies					
1	Tenaga Nasional Berhad (TNB)	10	Kumpulan Perangsang Selangor (KPS)	19	Kenanga Investment Bank
2	KPMG	11	Exim Bank	20	Kuala Lumpur Kepong Berhad (KLK)
3	Nippon Paint	12	Hong Leong Bank	21	Starbucks
4	Bank Rakyat	13	Kwantas Corporation Berhad	22	Panasonic
5	Pharmaniaga	14	Sunway Berhad	23	Poh Kong
6	RHB Bank	15	Titijaya Land Berhad	24	LeapEd Sevices Sdn. Bhd.
7	7-Eleven	16	Digi Telecommunications	25	YTL Corporation Berhad
8	Great Eastern	17	BH Petrol		
9	Top Gloves	18	FedEx		

2.2. Text pre-processing

Before the dataset is effectively analysed and mined, corpus documents must be noise-free. The phrase “corpus pre-processing” refers to the practise of deleting duplicate and less informative phrases from a corpus in order to establish a clean corpus. Data cleaning was performed by removing typographical error or validating and correcting values against a known list of entities. The pre-processing involves several processes to prepare the text data for qualitative analysis which are as follows:

- i) Normalization: the data should be normalized or standardized to bring all of the variables into proportion with one another. Non-numeric qualitative data should be converted to numeric quantitative data.
- ii) Tokenization: this stage divides the given text into smaller sections called sentences, and the phrases into smaller portions called tokens. White space or line breaks separate tokens.
- iii) Stop words removal: stop words are words that are not relevant to the desired analysis. Stop words, such as a, and are or do are eliminated as they are usually seen in the papers and do not provide any useful information.

- iv) Stemming: Stemming is the process of reducing words to their original root. The aim of stemming is to limit the diversity in text data by transforming words into their common form. For example, “contributes”, “contributing” and “contributed” were converted to “contribute”.
- v) Building corpus: Each document is represented in the corpus by a sequence of pairings. The first digit of the pair conveys that the numeric ID relates to a word, while the second digit expresses how frequently that word occurs. For instance [(1,1), ...] where “1” refers to the word “Donation” (for example) and “1” refers to the amount of times in which the word happens in the texts.

2.3. Document clustering

Document clustering involves grouping the documents together which are similar to each other. Clustering is an approach of unsupervised machine learning that can be performed on unlabeled data. In this study, k-means algorithms are used to establish a set of k clusters and assign each document to a certain cluster. Based on Zainal *et al.* [35], in 1967, Macqueen first introduced the k-means clustering technique, which was a simple and unsupervised learning clustering approach. K-means is a partitioning clustering technique that is used to automatically cluster texts in a corpus to ensure that documents in one cluster are more similar than documents in other clusters. The similarity between examples is calculated based distance measure as:

- Measure type: numerical measures
- Numerical measure: cosine similarity

Cosine similarity is often used in text analysis to measure document similarity [36]. It is measured by the cosine of the angle between two vectors which determines whether two vectors are pointing in roughly the same direction. All documents are assigned to their nearest cluster (nearest is defined by the measure type). Next, the centroids (determined by the position of the center) of the clusters are recalculated by averaging over all documents of one cluster. The previous steps are repeated for the new centroids until the centroids no longer move or max optimization steps are reached. Figure 1 shows the operators involved in RapidMiner studio to perform document clustering process.

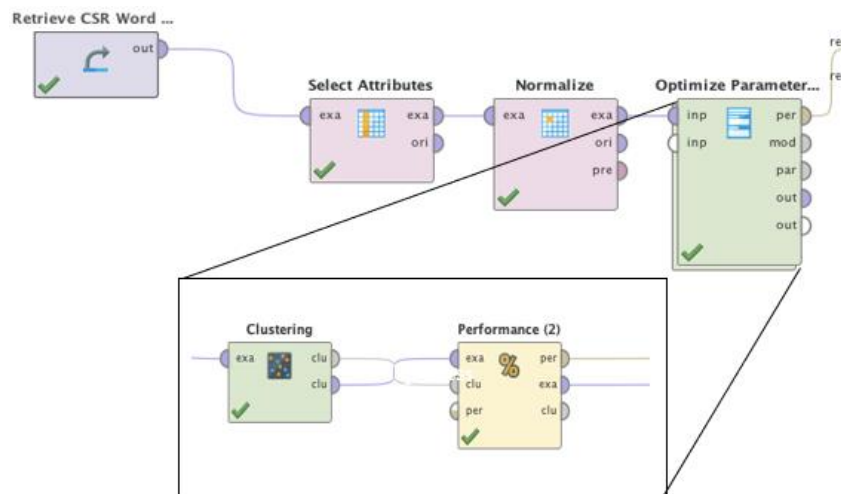


Figure 1. Document clustering process in RapidMiner

Based on Cheng *et al.* [37], the result of the document clusters is that a heterogeneous dataset is partitioned into several clusters with comparable members. However, the clusters created do not provide a description or characterization. Therefore, this study looks into topic modeling approach to further investigate the matter.

2.4. Topic modeling

Topic modeling seeks to locate a set of topics in a group of text documents; each topic is defined as a distribution across a set of words. This is performed by the application of statistical modelling techniques [33]. One of the most popular methods is LDA. LDA extends probabilistic latent semantic analysis (PLSA) by employing Dirichlet priors for document-specific subject mixtures, hence generating previously overlooked documents. LDA been a tremendous success in text mining due its excellent generalisation and extensibility [30]. LDA is based on the idea that each text document is composed of a large number of subjects, each of

which contains countless words. The input required by LDA is merely the text documents and the expected number of topics. LDA uses Gibbs Sampling for the application of the model.

LDA approach is used by randomly assigning a topic to each word in each document which involves 75 documents in this study. Two frequencies can be computed based on the word distribution of all subjects namely topic-document and word-topic. The popularity of each topic in the document is measured by the frequency counts calculated during initialization (topic frequency) and a Dirichlet-generated multinomial distribution over topics for each document. Meanwhile, the popularity of the word in each topic is measured by the frequency counts calculated during initialization (word frequency), and a Dirichlet-generated multinomial distribution over words for each topic. Then, reassign the word to the topic with the largest conditional probability. Lastly, iterate the process until the word to the topic assignments become significant in a stable state. In RapidMiner, operator extract topic from data (LDA) is used to execute the process.

3. RESULTS AND DISCUSSION

3.1. Text pre-processing

In this phase, the dataset was cleaned using the text-processing process which produced three numerical values for each document namely document length, token number and token length. This is derived from operators such as extract length, extract token number, and aggregate token length which are later used in the clustering process. The input and output of the text cleaning process are shown in Figure 2. From this process, wordlist, and word vectors for PCSR activities can be derived.

Row No.	TEXT	COMPANY	YEAR
1	Through Universiti Tenaga Nasional (UNITEN), we have been contributing towards the ...	TNB	2019
2	NET PROFIT ATTRIBUTABLE TO OWNERS OF THE COMPANY RM3,723.70 million. Joint ...	TNB	2018
3	NET PROFIT AT TRIBUTABLE TO OWNERS OF THE COMPANY RM6,904.0million. Total ...	TNB	2017
4	Legal entity for social enterprises In July 2020, the Dutch government announced a ne...	KPMG	2019
5	KPMG has committed itself worldwide to contributing to the 17 Sustainable Developme...	KPMG	2018
6	Jan Hommen Scholarship (founded in 2016) 8 students were awarded a EUR 2,500 sc...	KPMG	2017
7	In recent years, the three perspectives of Environment (E), Society (S), and Governance...	NIPPON PAINT	2019
8	Net income (100 million yen) 371. Contributions to local communities through table te...	NIPPON PAINT	2018
9	Net Income (billions of Yen) 36.01. In November 2016, we held a Global Quality Confe...	NIPPON PAINT	2017
10	Total Assets RM109.62 billion. To effect positive changes in society by providing islam...	BANK RAKYAT	2019
11	Total Assets RM106.89 billion. Profit after Taxation and Zakat (RM1.76 Billion). Contri...	BANK RAKYAT	2018
12	Total Assets RM105.45 billion. Benefiting needy communities through zakat contributi...	BANK RAKYAT	2017

(a)

Row No. ↑	text	COMPANY	YEAR	document_...	token_num...
1	univers tenaga nasion uniten contribut develop skill t...	TNB	2019	5514	881
2	net profit attribut owner compani million joint collab...	TNB	2018	3120	503
3	net profit tribut owner compani million total asset mi...	TNB	2017	4190	710
4	legal entiti social enterpris juli dutch govern announc...	KPMG	2019	1338	235
5	kpmg commit worldwid contribut sustain develop go...	KPMG	2018	676	112
6	jan hommen scholarship found student award eur s ...	KPMG	2017	3269	557
7	recent year perspect environ societi govern attract a...	NIPPON PAINT	2019	6450	1099
8	net incom million yen contribut local commun tabl te...	NIPPON PAINT	2018	4272	747
9	net incom billion yen novemb held global qualiti conf...	NIPPON PAINT	2017	3454	585
10	total asset billion effect posit chang societi islam fina...	BANK RAKYAT	2019	11056	1857
11	total asset billion profit taxat zakat billion contribut ...	BANK RAKYAT	2018	3944	654
12	total asset billion benefit needi commun zakat contri...	BANK RAKYAT	2017	2712	453

(b)

Figure 2. Data cleaning of (a) input process and (b) output process

3.2. K-means clustering

After performing text pre-processing, the word vector created was 7564. For clustering purpose, the attributes selected had numerical values such as document_length, token_length and token_number. This is because k-means requires a numeric measure to allow the algorithm in performing the distance measure.

Normalisation of data was then carried out to minimise the magnitude of the variable to avoid difficulties in the grouping or clustering process (refer Figure 3).

Row No.	text	document_...	token_num...	token_leng...
1	univers univ...	17407	1761	9.885
2	net net_prof...	9854	1005	9.805
3	net net_prof...	13271	1419	9.352
4	legal legal_e...	4239	469	9.038
5	kpmg kpmg...	2131	223	9.556
6	jan jan_hom...	10357	1113	9.305
7	recent recen...	20436	2197	9.302
8	net net_inco...	13555	1493	9.079
9	net net_inco...	10937	1169	9.356
10	total total_a...	35015	3713	9.430
11	total total_a...	12474	1307	9.544
12	total total_a...	8579	905	9.480
13	corpor corp...	15685	1635	9.593

(a)

Row No.	text	document_...	token_num...	token_leng...
1	univers univ...	0.527	0.437	2.139
2	net net_prof...	-0.169	-0.215	1.852
3	net net_prof...	0.146	0.142	0.225
4	legal legal_e...	-0.687	-0.678	-0.903
5	kpmg kpmg...	-0.882	-0.890	0.957
6	jan jan_hom...	-0.123	-0.122	0.057
7	recent recen...	0.807	0.813	0.043
8	net net_inco...	0.172	0.206	-0.757
9	net net_inco...	-0.069	-0.074	0.238
10	total total_a...	2.151	2.122	0.506
11	total total_a...	0.072	0.045	0.914
12	total total_a...	-0.287	-0.302	0.682
13	corpor corp...	0.368	0.328	1.091

(b)

Figure 3. The dataset condition, (a) before normalization process and (b) after normalization process

Based on the data that has been normalised in Figure 3, the study needs to determine the number of clusters, K to be selected. Optimize parameters operator was then used to find the optimal value of K which executes subprocess of clustering and performance operators. K-means clustering technique is adopted to minimize the distance within clusters while it maximizes the distance between clusters. The result is displayed in Figure 4. With high dimensional data, it can be hard to know what is the “best” number of clusters. Therefore, this study adopted “elbow method” of cluster selection where the value of K=7 was selected.

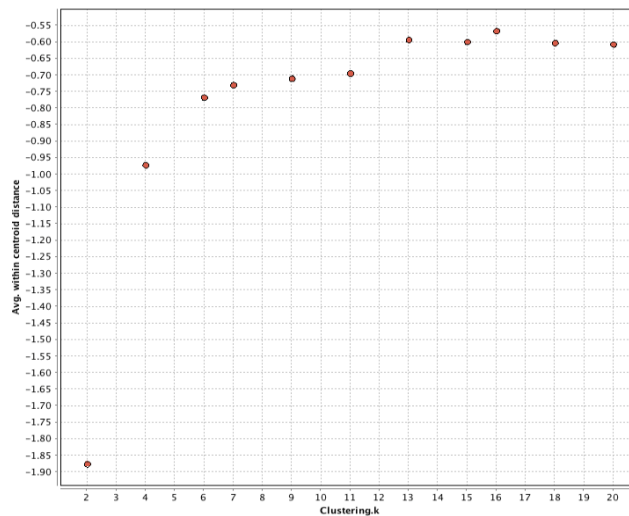


Figure 4. Scatter plot of k-means cluster

However, based on [32], it was reported that the mechanism in existing clustering models is lacking in terms of identifying local topics and global topics. The clustering algorithm on text data is a difficult process, and it is also a difficult task to obtain accurate results from clustering over text data. According to Lu and Castka [38], the unlabeled clusters that result from the clustering process can be characterized and explained using the topic modeling methodology. Therefore, integrating topic modeling might offer a global picture of the themes and their differences, while simultaneously enabling a comprehensive analysis of the keywords most closely linked to each topic.

3.3. LDA approach of topic modeling

In Malaysia, the most popular form of CSR activity is currently donations, grants, education support and sponsorships [39]. However, most recent categories are hard to find; therefore, it further justifies the

need in carrying out this study. The categories are used to further clarify the topic discovered through LDA approach particularly the Gibbs sampling technique. The same value of K selected in document clustering is also used in determining the number of topics in this part. The probabilities of each word in a “topic” are sorted in the descending order. When taking a close look, the results generated by LDA provide much more concrete and specific meaning. Each document is usually assumed to be generated by a few of the total number of possible topics. So, every word in every document is assumed to be attributable to one of the document’s topics.

Based on Table 2, it shows that the topic modelling has successfully found a set of topics in a group of text documents when applied to the entire dataset. LDA takes a term×document matrix as input and outputs the topic-word distribution using Gibbs Sampling approach. Then, the topic-word matrix contains the words that those topics can contain. Given the most common words in each topic, it is not hard to manually identify the names of each topic. Thus, from the five (5) top words in each topic, we can deduce the datasets related to the field of PCSR. As a result, the top five most frequent words are displayed in Table 2 which reflect the related concepts of each “topic”: Topic 0 is about community health and safety, Topic 1 is about education, Topic 2 is about employee welfare, Topic 3 is about social event, Topic 4 is about children development, Topic 5 is about local business development, and Topic 6 is about financial support.

Table 2. Top 5 most frequent words for each topic identified

Topic 0 Community health and safety	Topic 1 Education	Topic 2 Employee welfare	Topic 3 Social event	Topic 4 Children development	Topic 5 Local business development	Topic 6 Financial support
- Health	- Educational	- Employees	- Programme	- Group	- Million	- Bank
- Community	- Foundation	- Management	- Total	- Children	- Support	- Sunway
- Safety	- Students	- Work	- Organised	- Selangor	- Business	- Malaysia
- Provide	- Schools	- Program	- Students	- Staff	- Development	- Year
- Environment	- YTL	- Participated	- Malaysia	- Awareness	- Local	- Part

Based on the findings, the type of PCSR activities is plotted to each CSR company as shown in Figure 5 which represents topic distribution of documents (companies). The findings further explained the type of PCSR activities carried out by each company for the year 2017-2019. For example, TNB was involved with PCSR activities specifically on social event and local business development for 3-year period. Meanwhile, Nippon Paint was mostly involved with education, children development and local business development activities for that period of time. This insight gives companies an option whether to continue with their current volunteering works or to venture out into different kinds of PCSR activities. On the other hand, the government can use this information to work together with companies who are interested in specific PCSR activities to mitigate poverty eradication.

Company	PCSR Activities						
	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
TNB				X		X	
KPMG		X					
Nippon Paint			X		X	X	
Bank Rakyat				X			X
PharmaNiaga			X	X			
RHB Bank				X			X
7-Eleven	X			X	X		
Great Eastern				X			
Top Gloves	X						
KPS				X			
Exim Bank							X
Hong Leong Bank						X	X
Kwantas Corporation Berhad					X		
Sunway Berhad							X
Titijaya Land Berhad	X						
Digi Telecommunications		X				X	
BH Petrol						X	
FedEx						X	
Kenanga Investment Bank			X		X		
KLK		X	X				
Starbucks			X				
Panasonic	X		X				X
Poh Kong			X		X		
LeapEd Services Sdn Bhd		X					
YTL Corporation Berhad		X					

Figure 5. Categories of PCSR activities by each company for the year 2017-2019

4. CONCLUSION

This study is a preliminary investigation on the categorization of PCSR activities based on the annual report of selected CSR companies in Malaysia using an integrated approach of document clustering and topic modeling. Both approaches are a complementary technique which can minimize the limitation that each technique brings. As a result, the findings can categorize these activities into seven groups namely Community Health and Safety, Education, Employee Welfare, Social Event, Children Development, Local Business Development, and Financial Support. The findings can cluster PCSR activities that could assist companies and NGOs in managing funds distribution effectively based on the identified clusters that can be used in the government's intervention plan to solve poverty issue. In line with government programs, it is important to accurately predict allocations based on determined clusters because such knowledge may guide companies and practitioners in allocating the resources into those areas and populations where they are mostly needed. The findings from this study could provide a transparent distribution mechanism of corporate funds to benefit the community and the country, which can make a positive and significant impact on economic development both at the micro and macro levels. Moreover, the proposed hybrid model of both techniques does not only identify the hidden pattern of CSR activities that will be used as a strategic approach in planning, but the finding could also assist the government in strategizing and increasing participation of companies and NGOs in CSR activities for poverty mitigation programs.

There are multiple areas that can be further explored in the future. In the current implementation, k-means was used for the document clustering method, but in the future, fuzzy type clustering can be used as a proposed method since there is a chance that the data obtained are inconsistent or missing. Also, because of cluster limits overlap, some kind of patterns may be described in a single cluster group or a different group. K-means may not be successful in locating overlapping clusters as it also fails to solve issues such as incomplete external knowledge. Therefore, fuzzy type clustering can be used to overcome this drawback for the better output.





REFERENCES

- [1] X. T. T. Le and G. Teal, "A review of the development in defining corporate social responsibility," in *Science and Technology Development Journal*, vol. 14, no. 2, pp. 106-115, 2011, doi: 10.32508/stdj.v14i2.1935.
- [2] R. J. Baumgartner, "Managing corporate sustainability and CSR: A conceptual framework combining values, strategies, and instruments contributing to sustainable development," *Corporate Social Responsibility and Environmental Management*, vol. 21, no. 5, pp. 258-271, 2014, doi: 10.1002/csr.1336.
- [3] C. Flammer and J. Luo, "Corporate social responsibility as an employee governance tool: Evidence from a quasi-experiment," *Strategic Management Journal*, vol. 38, no. 2, pp. 163-183, 2017, doi: 10.1002/smj.2492.
- [4] L. S. Alrubaiee, S. Aladwan, M. H. A. Joma, W. M. Idris, and S. Khater, "Relationship between corporate social responsibility and marketing performance: The mediating effect of customer value and corporate image," *International Business Research*, vol. 10, no. 2, pp. 104-123, 2017, doi:10.5539/ibr.v10n2p104.
- [5] A. B. Carroll, "Managing ethically with global stakeholders: A present and future challenge," *Academy of Management Perspectives*, vol. 18, no. 2, pp. 114-120, 2004, doi: 10.5465/AME.2004.13836269.
- [6] S. Ilkhanizadeh and O. M. Karatepe, "An examination of the consequences of corporate social responsibility in the airline industry: Work engagement, career satisfaction, and voice behavior," *Journal of Air Transport Management*, vol. 59, pp. 8-17, 2017, doi: 10.1016/j.jairtraman.2016.11.002.
- [7] A. Amin-Chaudhry, "Corporate social responsibility—from a mere concept to an expected business practice," *Social Responsibility Journal*, vol. 12, no. 1, pp. 190-207, 2016, doi: 10.1108/SRJ-02-2015-0033.
- [8] N. A. Gardberg, S. C. Zyglidopoulos, P. C. Symeou, and D. H. Schepers, "The impact of corporate philanthropy on reputation for corporate social performance," *Business & Society*, vol. 58, no. 6, pp. 1177-1208, 2019, doi: 10.1177/0007650317694856.
- [9] S. Bose, J. Podder, and K. Biswas, "Philanthropic giving, market-based performance and institutional ownership: Evidence from an emerging economy," *The British Accounting Review*, vol. 49, no. 4, pp. 429- 444, 2017, doi: 10.1016/j.bar.2016.11.001.
- [10] I. E. Berger and V. Kanetkar, "Increasing environmental sensitivity via workplace experiences," *Journal of Public Policy & Marketing*, vol. 14, no. 2, pp. 205-215, 1995, doi: 10.1177/074391569501400203.
- [11] A. B. Carroll, "Carroll's pyramid of CSR: taking another look", *International Journal of Corporate Social Responsibility*, vol. 1, no. 1, pp. 1- 8, 2016, doi: 10.1186/s40991-016-0004-6.
- [12] A. B. Carroll and K. M. Shabana, "The business case for corporate social responsibility: A review of concepts, research and practice," *International Journal of Management Reviews*, vol. 12, no. 1, 2010, doi: 10.1111/j.1468-2370.2009.00275.x.
- [13] "Global sustainable development report, 2015 edition," New York, 30 June 2015, Accessed: January 3, 2022, [Online]. Available: <https://www.un.org/en/development/desa/publications/global-sustainable-development-report-2015-edition.html>
- [14] S. Imperatives "Report of the world commission on environment and development: our common future," 20 March 1987, Accessed: December 25, 2021, [Online]. Available: <https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf>
- [15] R. D. Medina-Muñoz and D. R. Medina-Muñoz, "Corporate social responsibility for poverty alleviation: An integrated research framework," *Business Ethics: A European Review*, vol. 29, no. 1, pp. 3-19, 2020, doi: 10.1111/beer.12248.
- [16] F. M. Musili, "Poverty-based classification of households using cluster analysis," M.S. thesis, Faculty of Science and Technology, University of Nairobi, US, 2020.
- [17] F.B. Ibrahim, "Strategy for Improving the Management of Major Flood Disasters in Kelantan (in Indonesia: Strategi Penambahbaikan Pengurusan Bencana Banjir Besar Di Kelantan)" Universiti Teknologi MARA, 2016.
- [18] M.S. Khalid, A. R. Anuar, and M. R. M. Jalil, "Disaster management system in Malaysia: study of flood cases (in Indonesia: Sistem pengurusan bencana di Malaysia: kajian kes banjir)", 2010.
- [19] N.W. Chan, "Impacts of disasters and disaster risk management in Malaysia: The case of floods." In *Resilience and recovery in Asian disasters*, pp. 239-265, Springer, Tokyo, 2015.





- [20] A. Frost, "A synthesis of knowledge management failure factors," *Recuperado el*, 22, pp. 1-22, 2014.
- [21] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: a simple yet principled alternative algorithm," *PLoS one*, vol. 11, no. 9, 2016, doi: 10.1371/journal.pone.0162259.
- [22] M. Abdul Rahman, N. S. Sani, R. Hamdan, Z. A. Othman, and A. A. Bakar, "A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group," *PLoS one*, vol. 16, no. 8, 2021, doi: 10.1371/journal.pone.0255312.
- [23] S. Das, A. Choudhary, and J. Harding, "An insight into corporate social responsibility reports: A text mining approach," *Conference: 3rd International Conference on Green Supply Chain, At: London*, 2016.
- [24] B. K. Chae and E. O. Park, "Corporate social responsibility (CSR): A survey of topics and trends using Twitter data and topic modeling", *Sustainability*, vol. 10, no. 7, 2018, doi: 10.3390/su10072231.
- [25] Z. Taran and B. Mirkin, "Exploring patterns of corporate social responsibility using a complementary k-means clustering criterion", *Business Research*, vol. 13, pp. 513-540, 2020, doi: 10.1007/s40685-019-00106-9.
- [26] A. Castellanos, C. M. Parra, and M. C. Tremblay, "Corporate social responsibility reports: Understanding topics via text mining," *Conference: Americas Conference on Information Systems (AMCIS) At: San Juan, Puerto Rico*, 2015.
- [27] R. K. Mishra, K. Saini, and S. Bagri, "Text document clustering on the basis of inter passage approach by using K-means," *International Conference on Computing, Communication & Automation*, 2015, pp. 110-113, doi: 10.1109/CCA.2015.7148354.
- [28] E. Sherkat, J. Velcin, and E. E. Milios, "Fast and simple deterministic seeding of k-means for text document clustering", *In International Conference of the Cross-language Evaluation Forum for European Languages*, vol. 11018, pp. 76-88, 2018, doi: 10.1007/978-3-319-98932-7_7.
- [29] F. J. M. Shamrat, Z. Tasnim, I. Mahmud, N. Jahan, and N. I. Nobel, "Application of k-means clustering algorithm to determine the density of demand of different kinds of jobs", *International Journal of Scientific & Technology Research*, vol. 9, no. 2, pp. 2550-2557, 2020.
- [30] F. Gurcan and N. E. Cagiltay, "Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling," in *IEEE Access*, vol. 7, pp. 82541-82552, 2019, doi: 10.1109/ACCESS.2019.2924075.
- [31] I. Sutherland and K. Kiatkawsin, "Determinants of guest experience in Airbnb: A topic modeling approach using LDA," *Sustainability*, vol. 12, no. 8, 2020, doi: 10.3390/su12083402.
- [32] M. Trupthi, S. Pabboju, and G. Narsimha, "Possibilistic fuzzy C-means topic modeling for twitter sentiment analysis," *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 3, pp. 100-108, 2018, doi: 10.22266/ijies2018.0630.11.
- [33] M. Alhawarat and M. Hegazi, "Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents," in *IEEE Access*, vol. 6, pp. 42740-42749, 2018, doi: 10.1109/ACCESS.2018.2852648.
- [34] BERNAMA, "41 corporations win sustainability and CSR Malaysia awards this year," Malaysia, 12 November 2020, Accessed: March 20, 2021, [Online]. Available: <https://themalaysianreserve.com/2020/11/12/41-corporations-win-sustainability-and-csr-malaysia-awards-this-year/>
- [35] K. Zainal, N. F. Sulaiman, and M. Z. Jali, "An analysis of various algorithms for text spam classification and clustering using RapidMiner and Weka," *International Journal of Computer Science and Information Security*, vol. 13, no. 3, pp. 66-74, 2015.
- [36] N. R. Mohammed and M. Mohammed, "Assessment of Twitter data clusters with cosine-based validation metrics using hybrid topic models," *Ingénierie des Systèmes d'Information*, vol. 25, no. 6, pp. 755-769, 2020.
- [37] X. Cheng *et al.*, "Topic modelling of ecology, environment and poverty nexus: An integrated framework," *Agriculture, Ecosystems & Environment*, vol. 267, pp. 1-14, 2018, doi: 10.1016/j.agee.2018.07.022.
- [38] P. Y. Shotorbani, F. Ameri, B. Kulvatunyou, and N. Ivezic, "A hybrid method for manufacturing text mining based on document clustering and topic modeling techniques," in *IFIP International Conference on Advances in Production Management Systems*, pp. 777-786, 2016, doi: 10.1007/978-3-319-51133-7_91.
- [39] J. Y. Lu and P. Castka, "Corporate social responsibility in Malaysia—experts' views and perspectives," *Corporate Social Responsibility and Environmental Management*, vol. 16, no. 3, pp. 146-154, 2009, doi: 10.1002/csr.184.

BIOGRAPHIES OF AUTHORS






Nik Siti Madihah Nik Mangsor     hold Bachelor of Science (Hons.) Computational Mathematics from Universiti Teknologi Mara (UiTM) Cawangan Kuala Terengganu, Malaysia. Currently she pursues her undergraduate studies in Master of Science (Computer Science) by research at UiTM Cawangan Kelantan. She is a Chief Executive Officer (CEO) of a charity company, Maab Barakah Resources in Machang, Kelantan. Her company interest in identifying the strategy to solve the poverty issue in Malaysia. She can be contacted at email: niksitimadidah@gmail.com.






Dr. Syerina Azlin Md Nasir     received her Ph.D. in Information Technology from Universiti Teknologi MARA (UiTM), Malaysia and her undergraduate studies at University of Salford, United Kingdom. She is a senior lecturer in Faculty of Computer and Mathematical Sciences at UiTM Cawangan Kelantan, Malaysia where she has been a faculty member since 2004. She has been engaged to research works such as conferences, workshops and become a member of Business Datalytics Research Group. She is a certified trainer in data analyst from RapidMiner and data integration from Talend. Her earlier publications are on database technology, ontology construction and mapping and actively involves in research, consultations and publications. The author's primary interest is on data mining, data analytics, text and web mining. She can be contacted at email: syerina@uitm.edu.my.






Dr. Wan Fairos Wan Yaacob    is a statistician by profession. She is a senior lecturer of Department of Statistics, Universiti Teknologi MARA Cawangan Kelantan, Malaysia and the Head of Business Analytics Department. She had a basic statistics degree from Universiti Teknologi MARA, Master of Science in Statistics from Universiti Kebangsaan Malaysia and PhD in Statistics from Universiti Teknologi MARA. Her area of research interest focuses on the development of panel count model estimator and evaluations on model parameters using monte-carlo simulation. She has published more than 35 articles and papers in various well-known journals and presented papers at conferences (H-Index 8, I-Index 7). She can be contacted at email: wnfairos@uitm.edu.my.



Dr. Zurina Ismail    is a Senior Lecturer at the Faculty of Business and Management, Universiti Teknologi MARA. She has been awarded with PhD in Marketing by Manchester Business School, The University of Manchester and her research is situated in the field of Strategic Marketing, with a special focus on Corporate Social Responsibility, Reputation Management and Business Model Innovation. She has actively involved in research and consultancy projects and has secured several agency and national grants. She can be contacted at zurinaismail@uitm.edu.my.



Dr. Shuzlina Abdul Rahman    is an Associate Professor of Information System at Universiti Teknologi MARA. She holds a PhD degree in Science and Systems Management with specialization in data mining and optimization. She was awarded as the recipient of the prestigious MIMOS 2012, the outstanding achievements of doctoral-level researchers in the areas of Information Communications and Technology (ICT). Her primary research interests involve the computational intelligence, data mining and optimization and intelligent data analytics. She can be contacted at email: shuzlina@uitm.edu.my.