Berkala Ilmu Perpustakaan dan Informasi, Vol. 15, No. 2, Desember 2019, Hal. 226-238
DOI: 10.22146/bip.36451
ISSN 1693-7740 (Print), ISSN 2477-0361 (Online)
Tersedia online di https://jurnal.ugm.ac.id/bip

Clustering topic groups of documents using K-Means algorithm: Australian Embassy Jakarta Media Releases 2006-2016

Wishnu Hardi[1], Wisnu Ananta Kusuma[2], Sulistyo Basuki[3]
[1] Overseas Collection and Metadata Management, National Library of Australia
[2] Program Studi Magister Teknologi Informasi untuk Perpustakaan, Institut Pertanian Bogor
[3] Sekolah Pascasarjana, Fakultas Ilmu Pengetahuan Budaya, Universitas Indonesia
e-mail: wishnu.hardi@dfat.gov.au

Naskah diterima: 23 Juni 2018, direvisi: 18 Juli 2019, disetujui: 24 Juli 2019

ABSTRAK

Pendahuluan. Kedutaan Australia menyimpan dokumen siaran media sebagai bentuk komunikasi resmi dalam konteks hubungan bilateral dengan Indonesia. Penelitian ini bertujuan melakukan klasterisasi informasi berbasis komputasi yang efisien dan sistematis berbasis penambangan teks sebagai upaya mengungkap pola, hubungan, dan struktur dokumen.
Metode penelitian. Algoritme K-Means digunakan sebagai metode klasterisasi nonhierarki yang mempartisi objek data ke dalam kelompok. Metode ini bekerja dengan miminimalkan variasi data yang berada dalam satu klaster dan memaksimalkan jarak antar klaster yang berbeda.
Analisis data. Analisis dilakukan terhadap 839 dokumen siaran media yang diterbitkan antara tahun 2006 sampai 2016. Proses analisis dimulai dari normalisasi teks dan transformasi data yang menghasilkan dataset baru untuk algoritme klasterisasi. Tahap evaluasi dilakukan melalui intepretasi data oleh pakar yang ditunjuk.
Hasil dan pembahasan. Proses ekstraksi informasi menghasilkan 57 istilah representatif dikelompokan ke dalam 3 klaster. Pakar menyimpulkan bahwa "hubungan antar masyarakat", "kerja sama ekonomi", dan "pembangunan kualitas hidup manusia" adalah konsep yang paling merepresentasikan topik dari dokumen siaran media Kedutaan Australia Jakarta tahun 2006 sampai 2016.
Kesimpulan dan saran. Penambangan teks dapat digunakan untuk klasterisasi topik dokumen. Dengan metode tersebut, klasterisasi dapat dilakukan lebih efisien dan sistematis karena proses analisis teks melalui sejumlah tahapan dengan parameter yang sudah ditentukan.

Keywords: *Penambangan teks; klasterisasi dokumen; algoritme K-Means, Cosine Similarity*

ABSTRACT

Introduction. *The Australian Embassy in Jakarta stores a wide array of media release document. Analyzing particular and vital patterns of the documents collection is imperative as it may result new insights and knowledge of significant topic groups of the documents.*
Methodology. *K-Means algorithm was used as a non-hierarchical clustering method which partitioning data objects into clusters. The method works through minimizing data variation within clusters and maximizing data variation between clusters.*
Data Analysis. *Of the documents issued between 2006 and 2016, 839 documents were examined in order to determine term frequencies and generate clusters. Evaluation was conducted by nominating an expert to validate the cluster result.*
Results and discussions. *The result showed that there were 57 meaningful terms grouped into 3 clusters. "People to people links", "economic cooperation", and "human development" were chosen to represent topics of the Australian Embassy Jakarta media releases from 2006 to 2016*
Conclusions. *Text mining can be used to cluster topic groups of documents. It provides a more systematic clustering process as the text analysis is conducted through a number of stages with specifically set parameters.*

Keywords: *Text mining; document clustering; K-Means algorithm, Cosine Similarity*

A. INTRODUCTION

Information technology has considerably changed the role of libraries in managing and disseminating information in the last three decades. From the 1970's to the 1980's all information managed by the libraries were in the formats of printed documents with limited accessibility. The Internet has played a significant role in shifting printed documents into digitalized ones, and the amount of these digital documents has rapidly risen until now. On the other hand, the capability of human to digest and process this huge amount of digitalized information remains constant.

Oracle corporation stated that nearly all organizations across the globe stored 80% of their unstructured data in their database, particularly those in the text formats. This trend tends to have significantly risen up to now (Mathew, 2012). Spire Technologies argued that 90% of institutional decisions were made on the basis of merely 20% of the structured data they have (Spire Technologies, 2016). This circumstance posed the essential inquiry on how every single organization may utilize 80% of their unstructured data effectively to ensure that the decisions made are on the basis of these data.

Text mining is a process of finding information among numerous collections of text and identifying interesting patterns of the information and their relationship within textual data (Feldman & Sanger, 2007). Text mining is a broad concept which reflects all approaches related to the analysis and processing of semi-structured and unstructured textual data. Seven text mining applications include information retrieval, classification, clustering, web mining, informational extraction, Natural Language Processing (NLP), and conceptual extraction. The commonality of these all approaches is how to convert textual data into numbers with strong algorithm basis which can be applied to large number of documents (Miner, Elder, & Nisbet, 2012).

In context of library and information sciences, text mining has existed for a long time, and it turned to existence a while ago as library and information scientists were developing textual analysis-based theories to measure information within a wide array of collections in the formats of printed documents. In terms of their main focus, these theories may be classified into two core groups. The first group focused on publication distribution through quantitative approach which includes text analysis, writer's productivity distribution, and word frequency's rank. On the other hand, the second group focused their work on citing analysis signified by so-called terms “impact factor” and “immediacy index” (Sulistyo-Basuki, 2014). Text mining currently tends to refer to information extraction processes, and its foundation is getting established and broadened due to the invention of numerous computation-based text analysis methods in the 2000's. Document collection is one essential element in text mining, and it is mainly aimed to organize and identify specified patterns of a huge pile of static and dynamic documents (Feldman & Sanger, 2007).

The Australian Embassy in Jakarta is one of the foreign government institutions storing a wide array of unstructured data in the formats of media release document. The importance of media release for one particular institution or organization was stipulated within the Vienna Convention on international relations. It clearly poses that one of the functions of the diplomatic institution is to represent the sending State in the receiving State for the purpose of furthering the development of commercial, economic, cultural and scientific relations between the sending State and the receiving State and otherwise promoting friendly relations between them in accordance with the provisions of the Convention (United Nations, 1961). Therefore, media releases are deemed strategically crucial as they are definitely diplomatic instrument to spread the sending State's foreign policies in the form of written document.

The number of the Australian Embassy Jakarta's media releases has fluctuated in the last ten years. The contents of the releases comprise official diplomatic communication on varied issues conveyed by the Australian government to Indonesian government and people in Indonesia in the context of bilateral relation between two countries. Analyzing particular and

vital patterns of the documents collection is imperative as it will result in new insights and knowledge which can be used as the guidance to make well-informed decisions and as the feedback to build diplomacy communications strategies. There has not been any single published study which specifically focused on text mining application to analyze the official documents released by embassies. Due to this fact, this study which highlights the text mining application to cluster topics of media releases issued by the Australian Embassy Jakarta from 2006 to 2016 is deemed crucially imperative.

B. LITERATURE REVIEW

1. Preprocessing

Data Preprocessing is the most elementary stage in text mining, at which raw data are transformed into more meaningful and understandable formats. It is common that textual data in the documents to be analyzed are not structured, consistent, and contain numerous noise; therefore, they require normalization for further processing. Preprocess stages include case folding, filtering, stopwords removal, and stemming (Solka, 2008).

2. Term Weighting

Term weighting is the assessment methodology to statistically evaluate level of importance of one particular term within a collection of documents. Term weighting requires two essential components: Term Frequency (TF) and Inverse Document Frequency (IDF). TF is statistical method to measure the occurrence frequency of one particular term within numerous documents, divided by total number of the term in the documents. Meanwhile, IDF aims to measure gap magnitude resulted by one specified term. This is due to the fact that the occurring term in the documents cannot be used to differentiate the documents for particular topics. Thus, the IDF of a rare term is high, whereas the IDF of a frequent term is likely to be low. The final TFIDF weight is the multiplication of both TF and IDF. For a term *i* in a document *j*, the weight *Wi,j* of term *i* in document *j* is given by:

$$W_{ij} = tf_{ij} \; x \log (D / df_i) \qquad (1)$$

here,

TFi,j is the number of occurrences of term *i* in document *j*, *DFi* is the number of documents containing the term *i*, and *D* is the total number of documents in the corpus (Salton, 1988).

3. K-Means Algorithm

K-Means algorithm was initially introduced by MacQueen in 1967, and it was used as a non-hierarchical data clustering approach which essentially grouping data objects into clusters. K-Means is a clustering process without any supervision whereby data objects are naturally deployed in one specified cluster without any initial knowledge or patterns to guide the clustering process. Within this approach, data with similar characters will be grouped into one cluster, whilst those with different patterns will be teamed up into other different clusters (Davis& Shaw, 2013).

4. Cosine Similarity

Cosine Similarity is an examining method of term commonality through measuring the angle between two vectors. In this method, a term is regarded as the entity having magnitude and direction in one high dimensional space. This method has extensively been used for clustering process including textual data clustering. By understanding distance between two terms, it is easy to identify commonality between the two terms. In particular, the similarity between the two terms can be identified from the minimum gap of distance between the two terms. Below is the equation used to identify similarity between the two terms:

$$Cosine\ (x, y) = \frac{\sum_{i=1}^{n} Xi.Yi}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \qquad (2)$$

where,

Xi and *Yi* are the components of the vector (features of the document, or TF-IDF values for each term of the document in this research), and *n* is the dimension of the vectors (Salton, 1988).

5. Within Cluster Sum-of-Square (WSS)

WSS is an evaluating method of intra-cluster variability. A cluster with a low WSS value has a higher cohesiveness value than that with a high WSS value. Each observation is allocated to the closest cluster, and the distance between an observation and cluster is calculated from the Cosine Similarity between the observation and the centroid. Each centroid will then be updated as the mean for observations in each cluster.The used equation is shown below:

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (X_{ji} - \overline{X}_{kj})^2 \qquad (3)$$

where,

S_k is the set of observations in the *k* cluster, and $\overline{X}_{kj}$ is the *j* variable of the cluster for the *k* cluster (Zhao, Xu, & Fränti, 2009).

6. Silhouette Coefficient

Silhouette Coefficient, introduced by Rousseeuw (1987), is a validating technique of cluster formation with graphical aid. It is mainly intended to determine the optimum number of clusters through measuring scale distance ratio among data objects into account. In addition, it is to identify cohesion and separation within particular Silhouette Coefficient value range [-1, 1] to determine data object similarity in clusters. Below is shown the equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (4)$$

For an individual point, *i,* calculate *a* = average distance of *i* to the points in its cluster, calculate *b* = min (average distance of *i* to points in another cluster). The algorithm starts with computing K-Means clustering algorithm for different values of *k* cluster. For each *k* cluster, calculate the average silhouette of observations. Plot the curve of average silhouette according to the number of clusters *k*. The location of the maximum is considered as the appropriate number of clusters (Rousseeuw, 1987).

7. Previous Studies

Previous text analysis studies using text mining have already been conducted in a wide range of fields. Zade, J., Bamnote, D., & Agrawal (2017) studied the foundation of document clustering process using text mining; Allahyari et al. (2017), on the other hand, conducted a study on fundamental techniques and aspects of text mining application in the fields of health and bio-medicine. In addition, Gurusamy, Kannan, & Prabhu (2017) did research on the grouping of social media users' behavior using clustering technique on the basis of K-Means algorithm. Wahid, D.H., & Azhari (2016), furthermore, analyzed texts to identify Twitter users' sentiments toward celebrity-related topics. Meanwhile, Prilianti, K.R., & Wijaya (2014) developed clustering application on the basis of K-Means algorithm to identify trends among the topics chosen by undergraduate students for their thesis. In addition, Lama (2013) utilized text mining as the primary instrument to cluster the main issues on electronic newspaper headlines with K-Means algorithm.

C. METHODOLOGY

1. Data collection

This research used media releases in English taken from the Australian Embassy Jakarta's official webpage: http://indonesia.embassy.gov.au/jakt/MediaRelease.html as the main source data. As official written communication of the embassy, media release is used to report specific and brief information about an event or other happenings and the it can be freely accessed and used by public in Indonesia. In this research, text analysis toward the document was merely conducted through the whole title and texts, excluding images, tables, and other illustrations.

2. Research design

This research was an explorative research in which the whole process and technique were solely on the basis of the previous studies. The tool used within the research was the R programming language version 3.4.4. One of the great strengths of R is the user's ability to add functions in statistical computing process. Clustering process involved in this research is illustrated in Figure 1.

Figure 1 shows clustering process applied in this research through modifying text mining application used by (Solka, 2008), in which data was initially collected and data characteristics were analyzed afterwards. The collected data was subsequently normalized through data preprocessing which comprises case folding, filtering, stop words removal, and stemming. Case folding is the process of converting uppercase letters to lowercase ones for the purposes of consistency and easy textual comparison in data preprocessing. Filtering is mainly intended to eliminate all numbers, punctuations, emails, websites, and any other non-alphabetical entities. Stop words removal, additionally, was the process to remove all existing stop words using list of English stop words.

The final stage in preprocessing is stemming, which is aimed to return all the terms back into their root. Preprocess data result was then transformed to numeric formats through term weighthing by computing TF/IDF value of each term, which were represented in Term Document Matrix (TDM) as data inputs for clustering algorithm. Data dimension reduction was also applied in this research for the purposes of computer efficiency memory and more convergent clustering results. K-Means clustering was done to obtain the optimal cluster number based on WSS and silhouette coefficient value. K-Means clustering process is done through minimizing data object variety in one particular cluster and optimizing data variety within different clusters. The following are clustering stages using K-Means method, 1) specify the number of clusters (k) to be created; 2) select randomly k objects from the dataset as the initial cluster centers or means; 3) assigns each observation to their closest centroid, based on the Cosine Similarity between the object and the centroid; 4) for each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a k cluster is a vector of length p containing the means of all variables for the observations in the k cluster; p is the number of variables; 5) iteratively minimize the total within sum-of-square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached

After the optimal cluster number was obtained, evaluation by experts was executed. Evaluation in this study was completed through interpreting process toward clustering outputs, which was conducted by experts. The interpreting process included two key points. Firstly, determining cluster labels as document topics. Secondly, explaining the argumentation for chosen labels. Experts involved in data interpretation must have the following requirements, 1) have academic qualifications; 2) have experience and competence on the concept and implementation of Australia's foreign policy to Indonesia; and 3) have institutional capacity to conduct data interpretation. The experts who were selected for this research were Mr. Wijaya Kusuma, MBA, senior researcher on Political and Economic Branch of the Australian Embassy Jakarta.

D. RESULTS AND DISCUSSIONS

1. Data collection

At the stage of data collection, media releases issued from 1 January 2006 to 31 December 2016 were downloaded from the Australian Embassy Jakarta's official website http://indonesia.embassy.gov.au/jakt/MediaRelease.html. There were 898 documents issued between 2006 to 2016, but merely 839 of them were successfully downloaded. The rest was not able to be downloaded because the file is unavailable in the server. All 839 documents with *html* extension were then converted to *csv* formats for further processing. Example of media release can be seen on Figure 2.

2. Preprocessing

Preprocessing in this research refers to text normalization process to obtain more consistent outputs. The preprocessing comprises case folding, filtering, stopwords removal, and stemming. Table 1 illustrates stages in data preprocessing.

Table 1 presents data preprocess stages. Stop words removal process utilized list of 571 stop words for English. Stemming stage, on the other hand, utilized affix-based Porter stemming

algorithm (Porter, 1980). The entire preprocess could reduce 12598 to 8250 terms, equal to 39% reduction of term count. This research finding was aligned with the finding of research, concluding that stop words removal process was able to reduce term count by 20-30% (Kannan, & Gurusamy, 2014). Dolamic & Savoy (2010) expressed the same finding that 571 stopwords was capable of deducting term count by 30-50%.

3. Term Weighting

TFIDF value weighting process was done through initially determining the values of both TF and IDF for every single term. Multiplication of both values resulted in the TFIDF value which was clearly outlined in the below Term Document Matrix (TDM) TFIDF in Figure 3.

4. Data Dimension Reduction

Data dimension reduction was completed to maintain the most significant terms, deemed as data variables determining trends of the document topics. The threshold was set at the value of 0.79, indicating that the terms whose frequency was lower than 21% would definitely be removed. Data dimension reduction process resulted in a final list of 57 terms for further processing. The value of 0.79 was obtained from examination process over a series of iteration which combined with reduction of sparsity percentage (data variable with 0 value in the matrix cells) and Silhouette coefficient value.

5. K-Means clustering

Clustering stage was accomplished based on the data in the TFIDF matrix which was developed before using K-Means algorithm. Determining the number of optimal clusters was done through examining the number of clusters *k=3* to *k=6.* This examination was part of the initial observation to attain the *Within Cluster Sum of Squares* (WSS) and Silhouette values as the primary parameters to finalize the number of clusters to be formed. Cluster with WSS value close to 0 indicated its low level of variability toward its intra-cluster. In other words, the lower the WSS value is, the higher level the commonality among cluster members is. Table 3 shows evaluating result toward the k number of clusters along with WSS and Silhouette coefficient values which had been identified already.

Table 3 illustrates that from the examination of clusters *k=3* to *k=6* the smallest WSS value belongs to cluster *k=3* (19.6%) and Silhoutte coefficient value which was closest to 1 belongs to cluster *k=3* (0.10). This implies that the number of cluster *k=3* was the most optimal number of clusters to be generated. Figure 4 shows Principal Component Analysis (PCA) plotting results of cluster *k = 3* using

Figure 4 highlights the clustering of 57 terms to 3 different clusters with each cluster comprising different number of members. Each data variable or term has different coordinates within two-dimensional space. The next step was to group all the terms into particular clusters based on their minimal distance to the centroid, and this process resulted in the terms membership for each cluster as shown in Table 2.

Table 2 summarizes the number of term items in each cluster. It is clear that cluster 1 had the highest number of term items (17); cluster 2 came second with 16 term items; and cluster 3 came next with 24 term items. These term items in all clusters were basically linguistic components which were further analyzed by appointed expert. Another consideration to take into account was the fact that cluster formation was highly based on frequency of term occurrence. Hence, it was likely that one particular term was less meaningful and essential to be an element of topic builder.

6. Cluster Analysis

In this research data was analyzed by appointed expert to identify specified labels in each cluster. They were assigned to interpret every single term item in all clusters and determine cluster label concepts representing topics of collections of analyzed respective documents. Next, they were inquired to provide the reasons for selecting such labels. Data interpretation results by the expert can be explained as follow:

Cluster 1

Expert provided label "People to people links" to cluster 1 as one of the main elements of the media release was to build mutual understanding and respect among the citizens of the two countries through innovation, science and technology, media and cultural exchange.. In further analysis, expert pointed out that Leaders noted the successful hosting of the 3rd Indonesia Australia Dialogue in Yogyakarta, in August 2016. The dialogue promotes mutual understanding between our two countries by facilitating productive and interactive discussions between participants from a broad range of fields, including business, science, education and media. Leaders acknowledged the role of civil societies in helping to shape bilateral relations and welcomed their recommendations as part of the solution to responding to challenges faced by the two countries. They encouraged the convening of the fourth Indonesia Australia Dialogue in Australia in 2017/2018.

Leaders supported the establishment of Indonesian language centers in Darwin, Brisbane and Sydney in addition to the existing centers in other parts of Australia. Such language centers aim to promote the study of Bahasa Indonesia and culture across Australia. Leaders agreed that the more Australians and Indonesians who can speak each other's languages, the more we can increase mutual understanding, respect and friendship between our peoples.

Leaders recognized that innovation, science and technology, media and cultural exchange are vital to promote mutual understanding and respect between the two societies. Therefore, they committed to increase those areas of cooperation. In this context, Leaders welcomed the success of the New Colombo Plan. Since 2014, over 2000 Australians have studied in Indonesia under the Plan. Innovation and science were highlighted as new areas for bilateral collaboration that were crucial to economic development. Leaders welcomed the inaugural Australia-Indonesia Science Symposium held in Canberra from 28 November to 1 December 2016 that brought together over 100 leading scientists from Australia and Indonesia to consider innovative research in health, marine science, agriculture, and big data.

Cluster 2

Cluster 2 was labelled with "Economic cooperation" due to shared motive between the two countries to promote investment in economy and international trade to get mutual benefits. This strategy was executed through the medium of inter-government partnership and supervision on economic reformation in the micro-economic and macro-economic sectors. Expert explained that as the two largest economies in the region, Australia and Indonesia have shared interests and a common future. Leaders noted trade links are strong but have not yet reached their full potential. They recognized that as close neighbors in the world's most dynamic region opportunities are on our doorstep. Leaders committed to intensify our efforts to achieve a high-quality Indonesia-Australia Comprehensive Economic Partnership Agreement (IA-CEPA) to transform our economic partnership.

The two Leaders' commitment on IA-CEPA reflects our shared determination to demonstrate regional leadership in support of trade to drive growth and prosperity. IA-CEPA will improve on commitments under our existing free trade agreement, the ASEAN-Australia-New Zealand Free Trade Agreement. It will create new openings for trade and investment to ensure the economic relationship can flourish into the future. Vision for IA-CEPA goes beyond a traditional free trade agreement. It will face the challenges and seize the opportunities of the current trading environment building on the special Indonesia–Australia relationship. Accordingly, IA-CEPA will deliver on trade and investment, and also drive cooperation and support greater partnership. Negotiations are progressing well, with the last round of talks held just last week. Leaders recommitted to concluding a comprehensive deal this year.

Australian economic governance investments support Indonesia to boost inclusive growth and to achieve mutual benefit

from international trade and investment. We provide technical assistance, including through government-to-government partnerships, which focus on Indonesia's priority economic reforms in areas such as financial sector supervision, budgeting, trade and competition, tax policy and administration and macroeconomic management. We are also working with Indonesia to tackle the underlying disincentives to investment in infrastructure, providing input on regulations and project planning, along with targeted technical assistance to make sure that Indonesia gets good results from its own spend on infrastructure development.

As nearly two thirds of Indonesia's poor live in rural areas, our aid program continues to focus on development of the agricultural sector. We are encouraging inclusive economic growth by attempting to influence how agricultural markets work for the poor, improving food security, raising agricultural productivity, and helping to boost farmer's incomes and employment by addressing constraints such as access to loans.

Cluster 3

The label "Human development" was devoted to cluster 3 as collection of terms in this cluster stressed the importance of the Australian Government's policies in complementing the Indonesian Government's efforts to improve its citizens' quality of life, especially in the areas of health (to anticipate harmful diseases and improve nutrients of mothers and children). Australia is supporting Indonesian efforts to help its people access better quality services, including in the poorer eastern regions of Indonesia.

The education program is focused on getting better education outcomes and supports Indonesian-led efforts to improve teacher quality and learning in schools. Australian government also supports Indonesia to trial innovative approaches for delivering improved education outcomes. Australia is targeting Indonesia's future leaders with PhD and Masters scholarships, and through short courses. On the other hand, Australia also encourages alumni links and people-to-people connections between Indonesians and Australians. In the health sector, Australia is working with Indonesia to improve human health and animal health systems to reduce the global threat posed by emerging infectious diseases. Australia is also providing support to improve nutrition for Indonesian women, children and newborns. In addition, the broader governance programs contribute to improved health and education service delivery. The disaster management program provides scientific and policy support to Indonesia to improve its preparedness and response systems and stands ready to assist in the event of a humanitarian disaster.

E. CONCLUSIONS

This study has evidently shown that text mining can be used to cluster topic groups of numerous press releases. This method has provided a more systematic clustering process as the text analysis is conducted through a number of stages with specifically set parameters. Furthermore, it presents that K-Means algorithm that can be utilized to identify cluster patterns which result in the terms with similar characteristics and features being grouped into one cluster. Cluster validating method is proved to be helpful to measure cohesion level, whilst cluster separating method can be used to determine the number of optimal clusters. The evaluation conducted by experts has strengthened the evidence that the results of topic clustering are in line with the ultimate goals of diplomacy of the Australian Embassy to Indonesia which focus on the elements of "people to people links", "economic cooperation", and "human development." These are the main goals of the Embassy's diplomacy which it tried to promote to the Indonesian government and Indonesian people. Further study shall be carried out with additional period of analyzing the documents to ensure extended coverage of the documents. Clustering approach utilized in the next study may be combined with hierarchical clustering approach. Clustering algorithm examination needs to be explored further to identify effectiveness of each utilized algorithm for clustering outputs, especially in the aspects of cluster cohesion and separation.

Any feedback is welcome providing that it is contributing to the improvement of this research.

ACKNOWLEDGMENT

I would like to thank Dr. Eng. Wisnu Ananta Kusuma and Prof. Dr. Sulistyo Basuki as supervisors. Thanks to Mr. Robin Brown in Australia and Ms. Kestrilia Rega Prilianti, M.Sc. who have provided invaluable feedback, Mr. Ilan Asqalani, M. Sc., Project Manager at the ASEAN Foundation and Mr. Wijaya Kusuma, MBA, senior researcher at Political and Economic Branch of the Australian Embassy Jakarta, as the evaluators.

REFERENCES

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. Retrieved December 2017 from https://arxiv.org/pdf/1707.02919.pdf

Davis, C.H., & Shaw, D. (2013). *Introduction to information science and technology*. Medford, N.J.: American Society for Information Society.

Dolamic, L., & Savoy, J. (2010). When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*, *61*(1), 200–203.

Feldman, R., & Sanger, J. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. New York: Cambridge University Press.

Gurusamy, V., Kannan, S., & Prabhu, J. R. (2017). Mining the attitude of social network users using K-Means clustering. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(5), 226–230.

Kannan, S., & Gurusamy, V. (2014). *Preprocessing techniques for text mining*. Paper presented at the Recent Trends and Research Issues in Computer Science (RTRICS) Conference, India. Retrieved from https://www.academia.edu/35015140/Preprocessing_Techniques_for_Text_Mining

Lama, P. (2013). *Clustering system based on text mining using the K-means algorithm: News headlines clustering*. Turku University of Applied Sciences. Retrieved November 2017 from http://www.theseus.fi/handle/10024/69505.

Mathew, S. (2012). Financial services data management: Big data technology in financial services. *Oracle White Paper*. Retrieved November 2017, from http://www.oracle.com /us/industries/financial-services/bigdata-in-fs-final-wp-1664665.pdf.

Miner, G. D., Elder, J., & Nisbet, R. A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Cambridge : Academic Press.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. Retrived from https://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf

Prilianti, K.R., & Wijaya, H. (2014). Aplikasi text mining untuk automasi penentuan tren topik skripsi dengan metode K-Means Clustering. *Jurnal Cybermatika*, *2*(1), 1–6.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(Nov.), 53–65.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Solka, J. L. (2008). Text data mining: Theory and methods. *Statistics Surveys*, *2*, 94–112.

Spire Technologies. (2016). Making sense of unstructured data with Spire. Retrived February 2018 from http://spiretechnologies.com/making-sense-unstructured-hr-data-spire/.

Sulistyo-Basuki. (2014). *Senarai pemikiran Sulistyo Basuki: Profesor pertama ilmu perpustakaan dan informasi di Indonesia.* Jakarta: Ikatan Sarjana Ilmu Perpustakaan dan Informasi Indonesia.

United Nations. (1961). Vienna convention on diplomatic relations. *International and Comparative Law Quarterly, 10*(3), 600-615.

Wahid, D.H., & Azhari, S. N. (2016). Peringkasan sentimen esktraktif di Twitter menggunakan hybrid TF-IDF dan Cosine Similarity. *Indonesian Journal of Computing and Cybernetics Systems*, *10*(2), 207–218.

Zade, J., Bamnote, D., & Agrawal, P. (2017). Text document clustering using K-Means algorithm with its analysis and implementation. *Imperial Journal of Interdisciplinary Research*, *3*(2), 1528–1531.

Zhao, Q., Xu, M., & Fränti, P. (2009). Sum-of-squares based cluster validity index and significance analysis. *Adaptive natural computing algorithms, 5495*(313-322).

PIGURE LIST

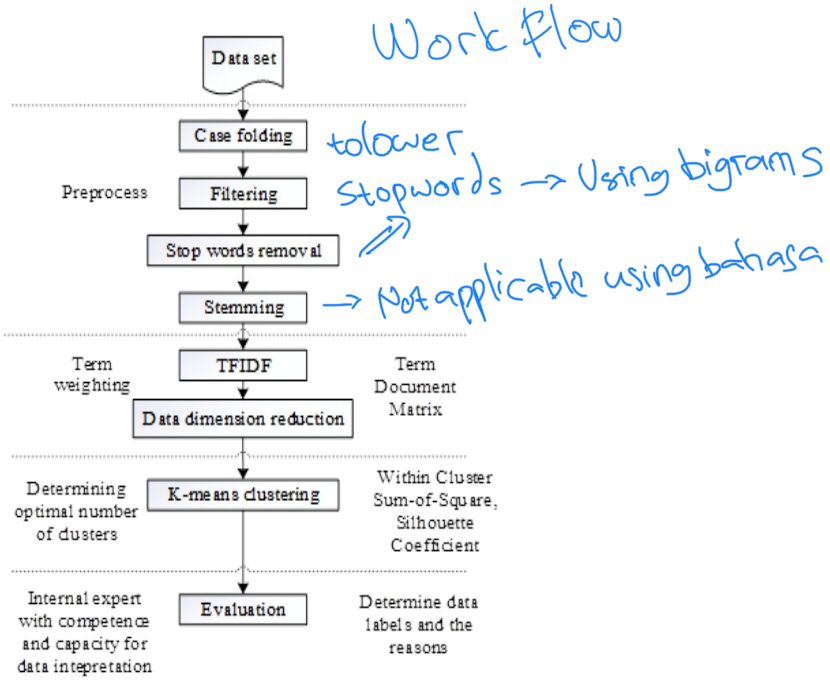

Figure 1. Clustering flow

Figure 2. Australian Embassy Jakarta media release

PIGURE LIST

```
<<DocumentTermMatrix (documents: 8, terms: 8)>>
Non-/sparse entries: 19/45
Sparsity           : 70%
Maximal term length: 7
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
    Terms
Docs      addit     advers      advis     affect      alloc     altern    announc     assist
   1 0.04528559 0.05017015 0.03526251 0.02319957 0.03203065 0.07422330 0.02781599 0.03830872
   2 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.03621739 0.02714573 0.01246187
   3 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.04137342
   4 0.00000000 0.00000000 0.00000000 0.00000000 0.03144827 0.00000000 0.00000000 0.00000000
   5 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.01620932 0.01488252
   6 0.00000000 0.00000000 0.00000000 0.02646712 0.00000000 0.00000000 0.00000000 0.00000000
   7 0.05166384 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.08740863
   8 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.01224066
```

Figure 3. Term Document Matrix (TFIDF)

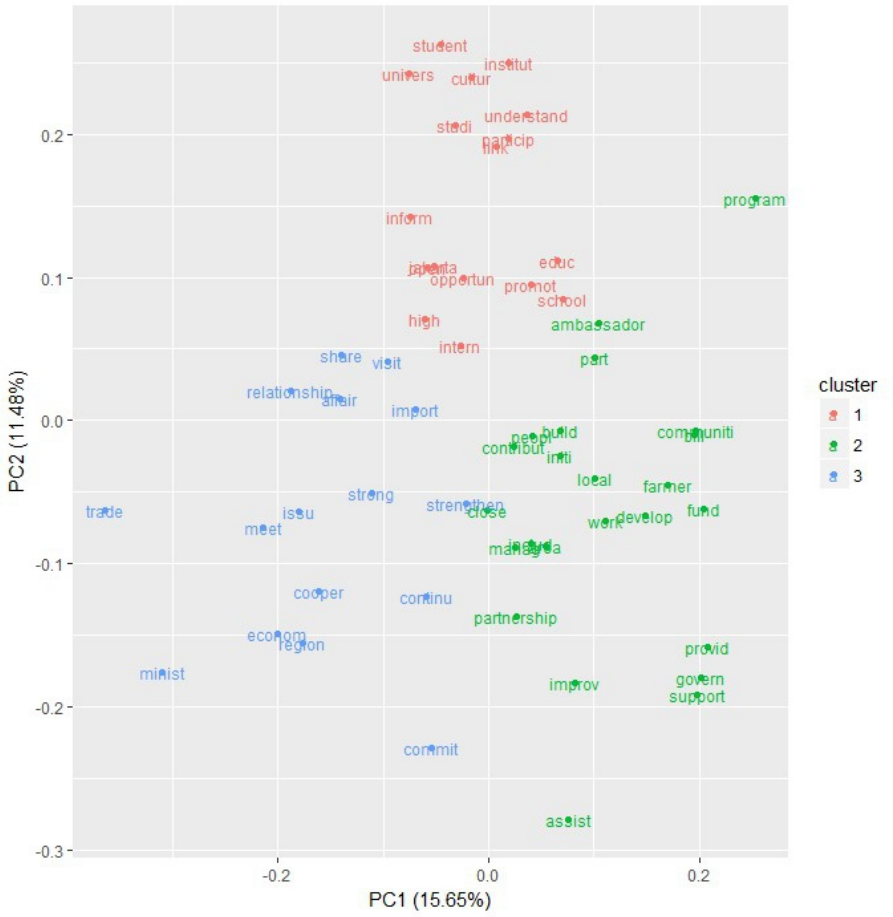

Figure 4. Principal Component Analysis (PCA) plotting of cluster k=3

TABLE LIST

Table 1. Data preprocessing stages

Preprocess	Outputs
Original text	Exhibition to Support Bali Rehabilitation Effort Today to commemorate the closure of the Bali Rehabilitation Fund (BRF)
Case folding	exhibition to support bali rehabilitation effort today to commemorate the closure of the bali rehabilitation fund (brf)
Filtering	exhibition to support bali rehabilitation effort today to commemorate the closure of the bali rehabilitation fund brf
Stop words removal	exhibition support bali rehabilitation effort commemorate closure bali rehabilitation fund brf
Stemming	exhibit support bali rehabilit effort commemor closur bali rehabilit fund brf

Table 2. Terms membership in cluster

Cluster	Number of membership	Term Items
1	17	{opportun, student, studi, high, school, institut, understand, jakarta, link, intern, cultur, educ, univers, particip, promot, open, inform}
2	16	{issu, relationship, continu, commit, trade, strengthen, cooper, meet , affair, econom, minist, visit, import, share, region, strong}
3	24	{assist, govern, close, contribut, part, build, improv, provid, local, communiti, includ, develop, initi, support, peopl, partnership, ambassador, farmer, program, fund, bill, work, mana g, area}