Social Media Mining: A Genetic Based Multiobjective Clustering Approach to Topic Modelling

Rayner Alfred, *Member, IAENG*, Loo Yew Jie, Joe Henry Obit, Yuto Lim, Haviluddin Haviluddin, Azreen Azman

***Abstract*—Social media mining is the process of collecting large datasets from user-generated content and extracting and analyzing social media interactions to recognize meaningful patterns in individual and social behavior. Everyday, more contents related to social media are generated by social media users (e.g., Facebook, Twitter). As the components of big data continue to expand, the task of extracting useful information becomes critical. Topic extraction refers to the process of extracting main topics from the pool of news feed and a typical method to perform topic extraction is through clustering. Clustering defines or organizes a group of patterns or objects into clusters, allows high-dimensional data to be presented in an apprehensive fashion to humans. Although effective, the performance of the k-means clustering algorithm depends heavily on the initial centroids and the number of clusters, k. Recently, several effective supervised and unsupervised machine learning methods have been developed in the domain of topics extraction. However, less works have been conducted in applying multiobjective based algorithm for topic extraction. Most of these algorithms are not optimized, even if they are, they are only optimized by using a single objective method and may underperform when solving real-world problems which are typically multi-objectives in nature. This paper investigates the effects of using a multiobjective genetic algorithm (MOGA) based clustering technique to cluster texts for topic extraction which is designed based on the structure and purity of the clusters in order to determine the optimal initial centroids and the number of clusters, k. Then, the mapping percentages between the predefined and produced clusters are used to assess the performance of the proposed algorithm. The best mapping percentage of 62.7 obtained using the proposed algorithm when k = 15 is obtained to outperform the performance of the generic k-means. The top five most representative words from each cluster are also extracted and validated by computing the number of tweets related to the predefined tags.**

***Index Terms*—Multi-Objectives, Genetic Algorithm, Clustering, Social Media Mining, Topics Extraction.**

I. Introduction

THERE are 3.8 billion people using social media today - more people than there were on the entire planet in 1971. In 2020, 3.8 billion is nearly half the world's population [1]. A massive amount of data is being published to social media sites every day, and all these unstructured data will be accumulated will not be processed in time. Thus, such continuing process leads researchers to investigate and produce more effective and efficient methods to extract, transform, load and analyze these massive amount of unstructured data referred as Big Data [2][3], which typically range from terabytes to petabytes. Big Data is defined with specific attributes that are called the five V's: volume, variety, velocity, value, and veracity [4][5][6]. Given the 5V's of Big Data, it is impossible for humans to assess and analyze such massive amount of data at once and it requires effective and efficient machine translation methods for computers to process and analyze these data using natural language processing methods, which is common to humans.

Topic extraction from microblogging sites has been the latest trend nowadays [7][8]. This process is important to help the community understand more on a topic and help them making informed decisions. Although most current topic extraction methods have its own unique capabilities, most of them are not optimized [9]. Since most topic extraction processes are done at the cluster level in which documents are grouped based on the contents, finding the optimized results of grouping or clustering results is another issue addressed by most researchers. Even if they are optimized, most of them are solved as a single objective problem. In order to improve the process of topic extraction, this work investigates the effects of using multiobjective optimization approach in extracting topics automatically, in which two main issues will be addressed, which are the optimized seeds for the initial centroids of the clusters and also the purity of the clusters. Several researches related to multiobjective clustering approach has been conducted recently [10][11][12][13]. However, most of them focused on internal validation measurements only. In this work, internal and external validation measurements are considered in optimizing the clustering process before topics extraction can be done. Therefore, this novel method propose a multiobjective genetic based clustering approach to handle topic extraction problem based on internal and external validation measurements.

In this work, tweets from microblogging social media sites such as Twitter are first extracted via its API and k-means clustering algorithm is then implemented to cluster them. Next, a Multi-objectives Genetic Algorithm (MOGA) based clustering technique is implemented to optimize the initial centroids of the clusters and also the purity of the clusters. Then, the performances of the proposed MOGA based clustering technique using different sets of fitness

Manuscript received March 26, 2020; revised September 15, 2020.

R. Alfred, Y.J. Loo and J.H. Obit are with the Knowledge Technology Research Unit, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia e-mail: ralfred@ums.edu.my, yewjie.loo@gmail.com, joehenry@ums.edu.my

Y. Lim is with School of Information Science,Security and Networks Area, Japan Advanced Institute of Science and Technology, Access 1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan e-mail: ylim@jaist.ac.jp

H. Haviluddin is with Department of Informatics, Universitas Mulawarman, Samarinda, Indonesia e-mail: haviluddin@unmul.ac.id

A. Azman is with the Department of Multimedia, Faculty of Computer Science and Information Technology, UPM, Malaysia e-mail: azreenazman@upm.edu.my

functions (single objective vs. multiobjective) are assessed. The mapping percentages between predefined and produced clusters are used to assess the performance of the proposed algorithm. Finally, the top five frequent words from each cluster are extracted to represent the topic of the cluster.

The rest of the paper follows, where Section 2 will highlight some related works on topics extraction. Section 3 will describe the experimental setup of this work. Section 4 discusses and analyses the results obtained. Section 5 will conclude the paper.

II. Related Works

In most conducted works on topics extraction methods, they have adopted the supervised and unsupervised machine learning approach. Earlier work involves applying the $TF * PDF$ algorithm approach [14] that assigns weight to each sentence in a corpus. Topic extraction is then performed using either classification or clustering in order to arrange the sentences chronologically. The novel concept of $TF * PDF$ is that a hot topic is normally discussed more frequent to allow equal importance from each newswire sources and channel them to the system in parallel [15]. In other works, Chen *et al.*'s [16][17] works aim to mitigate the information overload problem by focusing on important topics that appear with unusually high frequency during a specified timer period and typically contain several 'hot terms' that are the basis of topic extraction. Topic extraction is performed by mapping the distribution of the hot terms over time. Then, using multidimensional sentence vector, a clustering technique is applied to cluster the topics. 'Pervasiveness' and 'topicality' are the two vital properties in this method which can be used to improve the quality of the extraction process results. Nevertheless, not many works are found for non-English topic extraction (e.g., Malay language [18]) as it requires different set of resources such as Part of Speech (RPOS) Tagger for Malay [19], Named-Entity Recognition for Malay [20] and stemming for Malay language [21].

A classical clustering approach [22] considered a Probabilistic Cross-Lingual Latent Sentiment Analysis (PCLSA) model for topic extraction where the authors explain the main reason why existing topic models cannot be used for cross-lingual topic extraction. The topics are clustered using the General Expectation-Maximization (GEM) method. The work of Okamoto and Kikuchi [23] sees topic extraction carried out in a spatiotemporal theme pattern mining. Then, topics are clustered using a two-level hierarchical clustering. The first clustering method applies the agglomerative approach where the distance is measured using the Euclidean distance method. Then, the C-value technique is used to extract all the topic words for each topic cluster. Then, the subtopic clusters are extracted by using the second-level clustering. An optimized multi-layer ensemble framework has also been proposed and investigated for sentiment analysis [24]

Lastly, in Li *et al.*'s [25] work, a conversation tree is initially built, then using Conditional Random Fields (CRF), the leaders and followers across paths of conversation are detected to model microblogging topics. The detected leader/follower information is then incorporated as prior knowledge into the proposed topic model. This method is useful to model microblogging topics. Islam *et al.* proposed an improved online approach for clustering data stream into arbitrarily shaped with high accuracy, purity and noise sensitivity which can be applied in social media mining[26]. Online communities in a network can also be detected. For instance, Ajorlou *et al.* proposed a quality threshold clustering for detecting online communities in a network [27].

It is noticeable that recent topics extraction research are widely performed using machine learning methods and not many of them include the multiobjective optimization process as a research criterion in topic modelling or topic extraction. Even if they are optimized, most of them are solved as single objective problems which are unrealistic in real-world problems. Several researches related to multiobjective clustering approach has been conducted recently [10][11][12][13]. Mario *et al.* introduced and improved evolutionary approach to multiobjective clustering that applies Intracluster Variance (VAR) and Cluster Compactness (CNN)[11]. Intracluster Variance (VAR) reflects the compactness of the clusters while cluster connectedness reflecting the degree to which neighboring points are identified as members of the same cluster. In another work that proposed the application of Evolutionary Multiobjective clustering to patient stratification, five cluster validity indices were introduced that include compactness, separation, Calinski-Harabasz index, Davies-Bouldin index, and Dunn index [10]. However, all these cluster validity indices are considered as internal structure measurements. For instance, the cluster's compactness and separation's has been incorporated in Davies-Bouldin and Dunn indices. Both indices assume that better clustering means that clusters are compact and well-separated from other clusters. Similarly, Calinski-Harabasz index considers the overall within-cluster variance (equivalent to the total within sum of squares calculated above) and also the overall between-cluster variance. Many-objective fuzzy centroids clustering algorithm for categorical data has also been introduced [12]. Shang*et al.* utilized an artificial immune algorithm to address the multiobjective clustering problem and acquire a Pareto optimal solution set [13]. All of these works have used internal validation measurements only in constructing their evolutionary multiobjective clustering approach. In our work, an entropy index is introduced as an external validity measurement that considers the membership's purity of each cluster. Domain ontology can also be used for text pre-processing in order to improve the quality of the textual corpus being mined [28]. Several machine learning algorithms also have been used and investigated in identifying risk of cyberbullying from social network messages [29].

Genetic Algorithms (GAs) are designed based on the principle of survival of the fittest, by using cross-over and mutation mechanism [30]. GAs can be used to evolve the proper number of clusters and provide appropriate clustering [31][32][33]. By applying GAs, optimal solutions (individuals) can be obtained based on the predefined fitness functions used in the evolutionary process [30][34][35]. A GA-based clustering method called automatic genetic clustering has been developed for unknown k to automatically find the optimal number of clusters [36][37] using the Davies-Bouldin index to measure of the validity of clusters [38]. Being a population-based approach, GAs can be well suited to solve multiobjective optimization problems [37]. Multiobjective evolutionary algorithms (MOEAs) [39] have been

proven to bring promising solutions for such problems with effective search performance relative to single-objective clustering algorithms [40]. Optimization problem whether in single objective form or multiobjective form may be regarded as an issue or challenge that leave rooms to researchers in the field of topic extraction to look into. Therefore, it is essential to handle topic extraction problem using multobjective optimization approach so that more criteria can be taken into account. The work investigates the effect of using a Multiobjective Genetic Algorithm to determine the optimal initial centroids to partition the data points into clusters and to find better clustering solutions. Furthermore, internal and external validation measurements will be considered in constructing the evolutionary multiobjecrive clustering approach to handle topic extraction problem.

III. Methodology

There are three specific phases involved in this work which covers: Data Acquisition and Preprocess phase, Data Modelling phase, and Assessment phase. In the Data Acquisition and reprocess phase, a collection of 5000 tweets are collected and they are preprocessed and transformed into *term-document* matrix representation. In the Data Modelling phase, the solution of the problem represented by the chromosome is designed in which each chromosome is defined by the number of genes, g, that represents the number of documents collected in this work. In this phase, the crossover and mutation processes are also described to ensure that each solution or chromosome will have exactly k number of '1's. The size of the population is defined and the fitness function is defined based on two measurements which are Davis-Bouldin Index (DBI) and Variation of Information (VI) based on entropy. Next, in the Assessment phase, the quality of the clusters produced will be measured based on the value of the fitness functions used during the optimization process. The lower is the value of the fitness function the better is the quality of clusters produced. Several extended experiments will also be conducted in order to compute the percentages of mapping between the predefined clusters and the produced clusters. They are computed for all of the fitness functions used in this work in order to determine the best setting in finding the optimized clustering result. The visualization of terms extracted used for labelling the clustering will also be shown and discussed at the end of this paper. Fig. 1 illustrates all the phases involved in this paper.

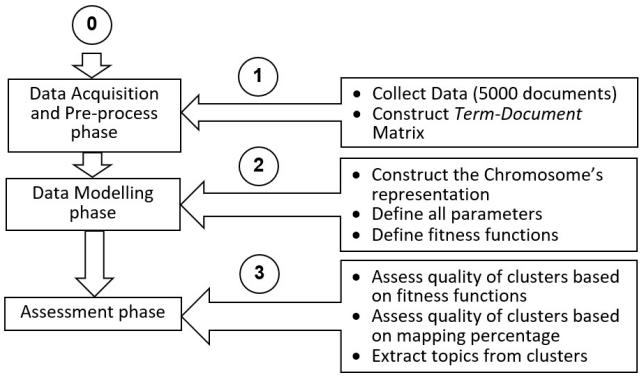

Fig. 1: Three phases involved in the research methodology

A. Data Acquisition and Preprocessing Phase

The datasets are acquired via Twitter API and stored as a 'list' data type where each element in the list represents an opinion/sentiment towards a targeted topic. The datasets collected are based on 10 trending topics found on Twitter with 500 tweets representing each topic. The overall number of documents, n, is 5000. The lists are then put together as a corpus for preprocessing purpose. Each topic will be treated as an initial cluster, resulting in a total of 10 initial clusters. This is also known as the predefined clusters, a, which will be used to compute the mapping quality based on Equation 9. Pre-processing involved parsing the collected tweets, removing stop words and transforming them into *term-document* matrix in which the weight of each term can be computed using the Term Frequency – Inverse Document Frequency (TF-IDF) formula outlined in Equation 3. The TF-IDF weighting scheme can be computed by the following:

$$tf_i = 1 + log(f_{i,j}) \tag{1}$$

$$idf_i = log(\frac{N}{n_i}) \tag{2}$$

$$tf - idf = w_{i,j} = 1 + log(f_{i,j}) * log(\frac{N}{n_i}) \tag{3}$$

where $f_{i,j}$ refers to the frequency of term i in document j, N refers to the total number of documents, n_i refers to the number of documents containing term i, and finally $w_{i,j}$ denotes the weight of term i in document j. The *term-document* matrix represents the weigtages assigned or computed for each unique term that exists in every document in the corpus and this *term-document* matrix can be used to perform the clustering task needed later in this paper [19][41].

B. Data Modelling Phase

This experiment adopts the binary encoding scheme [42]. Data points are applied in the dataset as the candidates for the cluster centers. The chromosome length is equal to the size of the data set. In this work, the length of the chromosome is 5000. The i-th gene of a chromosome corresponds to the i-th data point in the dataset. For a data point of index i to be the candidate for the center of a cluster, the allele of the corresponding i-th gene in the chromosome is set to "1"; otherwise "0". The number of clusters, denoted by k, is fixed within the range of 5 and 30. In generating the initial population, let P be the size of the population, P number of chromosomes will be generated and each chromosome will be represented by g, number of genes. In this work, g = 5000, since there are 5000 documents exist in the corpus. For initializing each chromosome in the P population, let g_k be distinct data points that will be randomly chosen from 1 to g, where g is number of genes. The gene corresponding to the index of each of these chosen data points is set to be "1"; while each of the remaining genes is set to be "0". For instance, given g = 16, g_k = 3 for chromosome and let 3 data points are randomly chosen from the dataset with indices 3, 10, 12, respectively, then the chromosome should have the following sequence of genes, 0010 0000 0101 0000.

There are two fitness functions that will be considered in this work. The first fitness function is called Davis-Bouldin

Index (DBI), has been selected for validating the clusters. DBI is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within the i-th cluster, S_i, is computed as

$$S_i = \frac{1}{\mid C_i \mid} \sum_{x \in C_i} D^2(z_i, x) \tag{4}$$

where $\mid C_i \mid$ denotes the number of data points belonging to cluster C_i, x denotes a single point that belongs to C_i, z_i denotes the centroid of the C_i cluster and $D^2(z_i, x)$ denotes the distance between the point x and z_i. Thus, Equation 4 compute the compactness of individual cluster i based on the average distance of all xs that belong to that cluster.

The distance between two cluster C_i and C_j is defined as $d_{i,j}$ in which the distance between the centers, z_i and z_j, is computed using Equation 5.

$$d_{i,j} = D^2(z_i, z_j) \tag{5}$$

Then, the DBI can be defined as follows;

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j, j \neq i} \frac{S_i + S_j}{d_{i,j}} \tag{6}$$

in which starting from cluster 1 to K, for every two clusters, i and j, the compactness of the clusters, S_i and S_j, and also the distance between two clusters, i and j, are computed. The worst scenario will be considered in which it will take the maximum value computed among all pairs of cluster 1 and the rest of the clusters. The process is repeated for the rest of the K-1 clusters.

The value of the DBI must be minimized in order to achieve proper clustering. In other words, the value of the DBI is smaller when the compactness of each cluster is high and the distance separating them is high. This is due to the fact that, a good clustering result depicts very compacted clusters and located far away from each other [43].

The second fitness function selected in this experiment is variation of information (VI). VI is a new measure introduced by an axiomatic view of clustering and is highly related to the mutual information, which measures the amount of information that is lost or gained in changing from the class set to the cluster set [44]. VI is the enhanced version of the entropy measure in which a lower VI value implies a higher clustering quality and it can be computed as follows:

$$VI = -\sum_{i=1}^{a} p_i \log_2 p_i \tag{7}$$

where a denotes the number of classes and p_i refers to the probability of item i in that particular group or cluster.

Reproduction – In this work, a modified reproduction operator is applied by adding a so-called "winner replacing" step prior to the roulette wheel

Crossover – The crossover operation is performed each time on a single gene position in some other proposed algorithms. There is a possibility to produce a total number of clusters smaller or bigger than the predefined number of clusters, k, which is ranging between 5 and 30. This might leads to unreasonable offspring and need to be repaired for many generations. In order to overcome this situation, a different approach of mutation is therefore implemented.

Mutation – The conventional mutation operator is performed on a gene-by-gene basis. Provided with the rate of the mutation, each gene in all chromosomes in the whole population undergoes mutation. In order to ensure the number of '1' is always equal to the predefined number of k, the mutation operator is altered by first finding the number of '1', n, produced in the final solution and then compare it with the number of predefined number of k. If n is greater than the predefined number of k, then $n-k$ number of bits containing '1'will be chosen at random and then mutated to become '0'. If n is smaller than the predefined number of k, then $k-n$ number of bits containing '0' will be chosen at random and then mutated to become '1'.

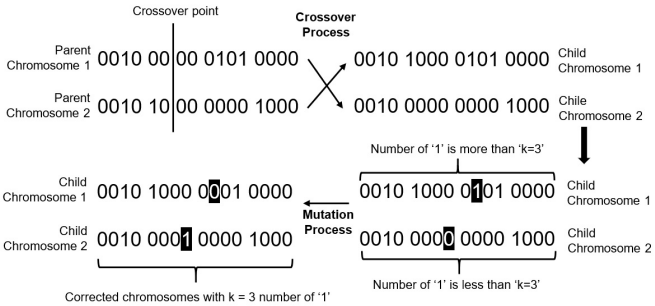

Fig. 2: Crossover and Mutation Processes

In Fig. 2, the number of clusters presented in the parent chromosomes, chromosome 1 and 2, are k = 3, but after the crossover process, the number of '1' that exists in the child chromosomes, Child 1 and 2, are 4 and 2 respectively. In child chromosome 1, since n is greater than the predefined number of k, then $n-k$ number of bits containing '1' will be chosen at random and then mutated to become '0'. On the other hand, in child chromosome 2 since n is less than the predefined number of k, then $k-n$ number of bits containing '0' will be chosen at random and then mutated to become '1'.

After the crossover and mutation processes, by using the predefined number of k, the fitness function that consists of DBI or VI is then optimized separately using a Genetic Algorithm (GA) as a single objective problem (SOGA). Next, the experiment is then carried out in a way where both fitness function (DBI and VI) are optimized simultaneously using the weighted sum approach where weight w_i is assigned to each normalized objective function $z_i^{'}(x)$ so that the problem can be converted to a multi-objectives problem (MOGA) with a scalar objective function as follows:

$$Min_z = w_1 z_1^{'}(x) + w_2 z_2^{'}(x) + \cdots + w_k z_k^{'}(x) \tag{8}$$

C. Assessment Phase

Genetic algorithms are stochastic in nature which implies that they will not give identical results every time they run [45]. In order to address this issue, the dataset and fitness function is run five times using the same parameters. For each set of the outcome obtained using the same parameter, the mean, variance and the minimum fitness values of the data are computed and recorded. Mapping percentage needed to be calculated from bidirections in order to make sure that both clustering processes are parallel [41]. The formula

used for mapping, $M(a,d)$ the clusters between predefined clusters, a, and the produced clusters, d is as followed,

$$M(a,d) = \frac{\sum_{1\leq i\leq j, 1\leq k\leq l} \frac{(\frac{|C_i(a)\cap C_j(d)|}{|C_i(a)|} + \frac{|C_i(a)\cap C_j(d)|}{|C_j(d)|})}{2}}{m} \quad (9)$$

where $\mid C_i(a) \mid$ is the number of documents in the predefined cluster and $\mid C_j(d) \mid$ is the number of documents in the produced clusters. Then, $\mid C_i(a) \cap C_j(d) \mid$ refers to the number of documents in the predefined cluster that can be mapped to produced documents or vice versa. m refers to the maximum number of possible pairings between predefined clusters and produced clusters, $l \times k$, in which l refers to the number of predefined clusters while k refers to the number of produced clusters. A higher value of mapping percentage would indicate that the predefined and produced clusters are more parallel to each other. For instance, Fig. 3 shows

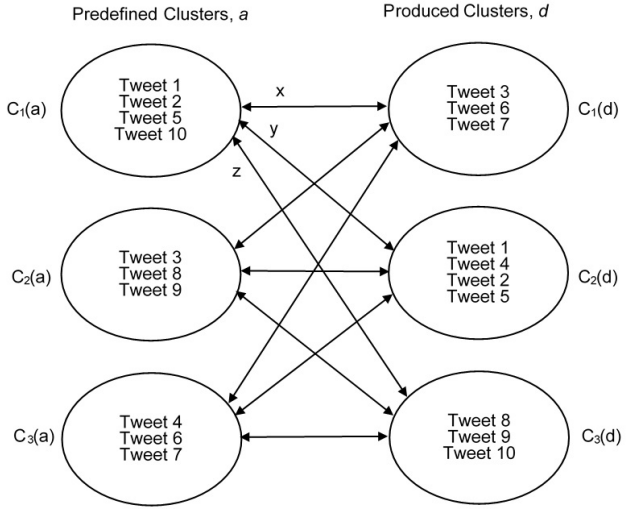

Fig. 3: Mapping clustering results between predefined cluster and produced cluster.

the mapping clustering between predefined clusters, a and produced clusters, d. Based on this diagram, l is 3 and k is 3, then $m = l \times k = 9$.. Then, $x = M(a_1, d_1) = \frac{|C_1(a)\cap C_1(d)|}{2}$ will give $\frac{\frac{0}{4}+\frac{0}{4}}{2} = 0$, $y = M(a_1, d_2) = \frac{|C_1(a)\cap C_2(d)|}{2}$ will give $\frac{\frac{3}{4}+\frac{3}{4}}{2} = 0.75$, and $z = M(a_1, d_3) = \frac{|C_1(a)\cap C_3(d)|}{2}$ will give $\frac{\frac{1}{4}+\frac{1}{4}}{2} = 0.25$. The overall mapping assessment can be computed by evaluating the mapping percentage according to Equation 9.

The proposed GA-based clustering algorithm involves several parameters that include the number of clusters, k, probabilities of crossover, P_c, population size, P_s and number of generations, G_n. The mutation rate is fixed at 1 to ensure the number of '1' equals the predefined number of clusters, k. The GA is first applied to find the best solution by optimizing the fitness functions of DBI and VI separately using different sets of parameters which includes, k = {5, 10, 15, 20, 25, 30}, Probability of crossover, P_c=(0.8, 0.5), single-point crossover, population size, P_s = 100, number of generations, G_n = 100. Then, a multiobjective GA is applied again to optimize the fitness functions of Davies-Bouldin Index (DBI) and Variation of Information (VI), using a weighted-sum approach, where different weights are assigned to each fitness function. The Multiobjective GA is carried out using different sets of parameters which includes, fitness functions (MO) as follow,

$$MO_0 = 0.9 \times DBI + 0.1 \times VI \quad (10)$$

$$MO_1 = 0.7 \times DBI + 0.3 \times VI \quad (11)$$

$$MO_2 = 0.5 \times (DBI + VI) \quad (12)$$

$$MO_3 = 0.3 \times DBI + 0.7 \times VI \quad (13)$$

$$MO_4 = 0.1 \times DBI + 0.9 \times VI \quad (14)$$

and k = {5, 10, 15, 20, 25, 30}, P_c=(0.8, 0.5), crossover type = single-point, population size, P_s = 100, number of generations, G_n = 100. In order to eliminate the stochastic nature of SOGA and MOGA, each test is ran five times using different seeds across selected sets of parameters. In this work, the optimization is performed by minimizing the value obtained for each MO_i stated in Equation 10, 11, 12, 13, 14, since the lower the values for both DBI and VI, the better is the result of the clustering process.

Several extended experiments are also conducted with the following parameters' values, that include k = 5, 10, 15, pair of fitness functions for both VI and DBI = {(1.0, 0.0), (0.9, 0.1), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7), (0.1, 0.9), (0.0, 1.0)}, P_c=0.5, population size = 100, number of generations = 100. The results of the extended experiments are tabulated in Table V. The purpose of conducting these extended experiments is to compare the mapping percentages obtained using Equation 9, when using different parameters mentioned earlier for both SOGA and MOGA methods.

IV. Results and Discussion

Results for SOGA and MOGA are tabulated in Table I and IV, where the best fitness is recorded. Results of extended SOGA and MOGA are tabulated in Fig. 6, where the best fitness is also recorded. Based on the results shown in Table I, it is noticeable that putting weight which is biased more towards VI normally yields the best solution when k = 10, 15, 20, 25 and 30, while DBI yields best solution when k = 5. This occurrence is acceptable and explainable since the default number of selected topics to represent the initial clusters is set at 10. It is generally obvious that putting weight which is biased more towards VI would result the clustering process to focus more on the label of the documents rather than the structure of the clusters. The five different weighted-sum approaches are:

- MO_0: $0.9 \times DBI + 0.1 \times VI$
  - Weight is biased more towards DBI
- MO_1: $0.7 \times DBI + 0.3 \times VI$
  - Weight is biased towards DBI
- MO_2: $0.5 \times (DBI + VI)$
  - Weights are equally assigned
- MO_3: $0.3 \times DBI + 0.7 \times VI$
  - Weight is biased towards VI
- MO_4: $0.1 \times DBI + 0.9 \times VI$
  - Weight is biased more towards VI

Based on the t-test results shown in Table II, the p-value is less than the alpha level: $p < 0.05$ for VI and MO_2, MO_3,

TABLE I: The comparison of fitness value between SOGA and MOGA when P_c=0.8 and P_c=0.5

		SOGA		MOGA				
k	P_c	DBI	VI	MO_0	MO_1	MO_2	MO_3	MO_4
5	0.8	**1.8**	2.3	**1.8**	1.9	2.4	2.7	2.4
	0.5	**1.9**	2.3	**1.9**	2.4	2.3	2.8	2.4
10	0.8	2.5	**2.2**	2.5	2.7	2.7	2.6	**2.5**
	0.5	2.6	**2.1**	2.4	2.4	2.6	2.7	**2.4**
15	0.8	2.5	**2.1**	2.5	**2.4**	2.6	2.6	2.5
	0.5	2.3	**2.1**	2.4	2.5	2.6	2.5	**2.4**
20	0.8	2.4	**2.2**	2.4	2.5	2.5	2.4	**2.3**
	0.5	2.4	**2.2**	2.4	2.4	2.6	2.5	**2.4**
25	0.8	**2.3**	**2.3**	**2.4**	2.5	2.6	2.5	**2.4**
	0.5	2.5	**2.3**	2.5	2.6	2.7	2.5	**2.4**
30	0.8	2.4	**2.3**	2.4	2.5	2.5	2.5	**2.4**
	0.5	2.4	**2.3**	2.5	2.6	2.6	2.5	**2.4**

TABLE II: The comparison of t-test results (p-value) between SOGA and MOGA when P_c=0.8

	SOGA		MOGA				
	DBI	VI	MO_0	MO_1	MO_2	MO_3	MO_4
DBI	-	0.48	0.92	0.53	0.07	0.07	0.39
VI	0.48	-	0.40	0.14	0.000	0.000	0.002
MO_0	0.92	0.40	-	0.60	0.09	0.09	0.48
MO_1	0.53	0.14	0.60	-	0.29	0.29	1.00
MO_2	0.07	0.000	0.09	0.29	-	1.00	0.029
MO_3	0.07	0.000	0.09	0.29	1.00	-	0.029
MO_4	0.39	0.002	0.48	1.00	0.03	0.03	-

MO_4. As a result, we can reject the null hypothesis that there is no difference between VI and MO_2, MO_3, MO_4. Similarly, we can also reject the null hypothesis that there is no difference between MO_2 and MO_4, MO_3 and MO_4. In other words, applying the single objective VI can produce significant improvement compared to applying multiple objectives. Among the multi-objective methods, MO_4 can be used to produce significant improvement compared to MO_2 and MO_3. Similar patterns can be observed when P_c=0.5 as shown in Table III.

It is noticeable that smaller number of k normally yields a better solution in the above mentioned weighted-sum approach. By comparing the six different weighted-sum approaches, it is found that the best solution (minimum value) usually occurs when weight is biased towards VI (e.g., MO_4) and when weight is biased towards DBI (e.g., MO_0), but never when weights are equally assigned (e.g., MO_2). When weight is biased towards VI (e.g., MO_4), the performance of minimization is slightly better than the case when weight is biased towards DBI (e.g., MO_0).

Next, the mapping function is applied to evaluate the performance of the produced clusters. A higher value of mapping percentage would indicate that the predefined and produced clusters are more parallel to each other. Fig. 4, 5 and 6 illustrate the Pareto analysis for different number of k and P_c = 0.5, while their results are tabulated in Table V. This analysis provides the trade-off between DBI, VI, and the mapping percentages produced when the weights vary. These graphs will provide useful insights for decision makers to cast their decisions. Based on the Pareto analysis, it is clear that optimizing SOGA with VI yields better solutions relative to DBI. In MOGA, when VI is assigned a greater

TABLE III: The comparison of t-test results (p-value) between SOGA and MOGA when P_c=0.5

	SOGA		MOGA				
	DBI	VI	MO_0	MO_1	MO_2	MO_3	MO_4
DBI	-	0.24	1.00	0.24	0.09	0.07	0.63
VI	0.48	-	0.21	0.001	0.000	0.000	0.001
MO_0	0.92	0.40		0.21	0.07	0.05	0.60
MO_1	0.53	0.14	0.21	-	0.25	0.17	0.06
MO_2	0.07	0.000	0.07	0.25		0.83	0.014
MO_3	0.07	0.000	0.05	0.17	0.83	-	0.007
MO_4	0.39	0.002	0.60	0.06	0.014	0.007	-

TABLE IV: The comparison of fitness values between SOGA and MOGA when averaged across five runs with P_c=0.5

k	SOGA	MOGA
5	DBI: 1.9	MO_0: 1.9
10	VI: 2.1	MO_0: 2.4, MO_4: 2.4
15	VI: 2.1	MO_4: 2.4
20	VI: 2.2	MO_4: 2.4
25	VI: 2.3	MO_4: 2.4
30	VI: 2.3	MO_4: 2.4

weight, the solutions produced are also better than that when the weight is equally distributed among DBI and VI, and when the weight is biased towards DBI. This is due to the fitness function of VI, which tries to search for better purity and membership of the clusters. The mapping percentages are then computed again using the solutions produced in the extended experiments on SOGA, MOGA and also using the generic and un-optimized k-means clustering. The mapping percentages are tabulated in Fig. 7. Through the mapping percentages obtained, it is clear that SOGA and MOGA perform better than generic k-means clustering algorithm. This shows the optimal mapping percentages is found when using SOGA using VI as the fitness function with P_c=0.5 at k=5.

TABLE V: Mapping Percentage obtained with different set of weights assign to fitness function when k=5, 10 and 15 and Pc=0.5

k	DBI	VI	Mapping %
5	1.0	0.0	21.13
	0.9	0.1	21.14
	0.7	0.3	21.14
	0.5	0.5	21.02
	0.3	0.7	40.08
	0.1	0.9	**41.12**
	0.0	1.0	31.04
10	1.0	0.0	42.23
	0.9	0.1	41.24
	0.7	0.3	35.24
	0.5	0.5	39.67
	0.3	0.7	46.73
	0.1	0.9	**70.45**
	0.0	1.0	69.67
15	1.0	0.0	40.48
	0.9	0.1	43.23
	0.7	0.3	47.08
	0.5	0.5	45.85
	0.3	0.7	61.32
	0.1	0.9	70.12
	0.0	1.0	**73.09**

Observing the trends illustrated in Fig. 4, Fig. 5 and Fig. 6, it is obvious that higher mapping percentage can be obtained by putting weight which is biased more towards VI.

Topics extraction are finally performed where top five most

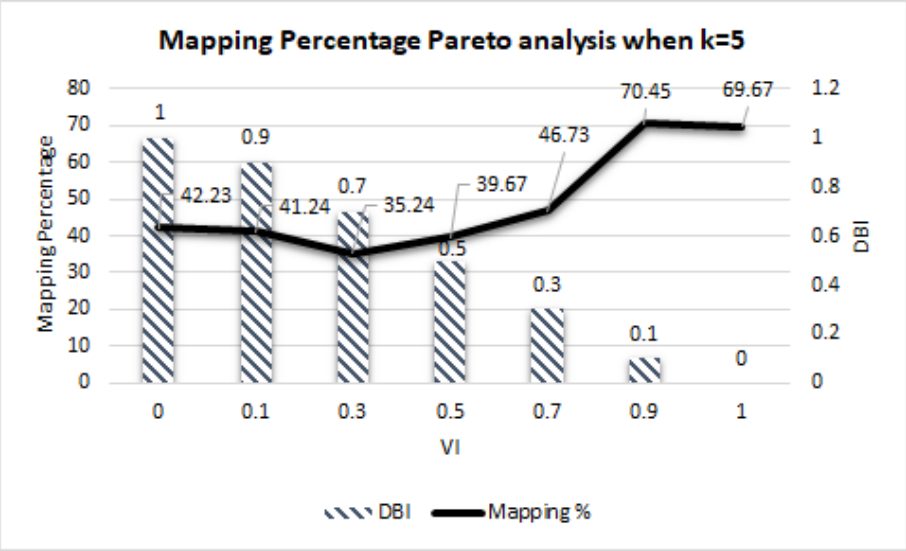

Fig. 4: Mapping Percentage Pareto analysis when k=5 and P_c=0.5

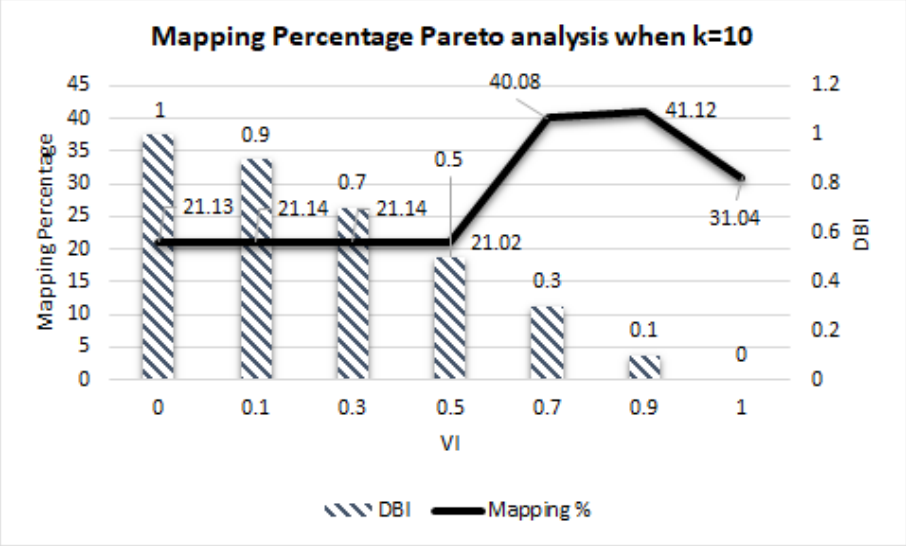

Fig. 5: Mapping Percentage Pareto analysis when k=10 and P_c=0.5

representative words from each cluster are extracted and validated by computing the number of tweets related to the predefined tags. Since the mapping percentage only consider the maximum pairing between the predefined clusters and produced clusters, therefore topics from the overall best mapping percentage cluster will only be extracted. The optimal solutions for each tested method and for each number of k are tabulated in Table VI to VIII respectively, where the top five most representative words and number of predefined tags are showed.

In Table VI, SOGA and MOGA based k-means produced a cluster size of 450 each. 449 of them belong to the pre-defined tag of "Great Barrier Reef" and with one irrelevant tag detected respectively. Meanwhile, the generic k-means produced a cluster size of 417. 414 of them belong to the predefined tag of "Note 7" with three irrelevant tags detected. In Table VII, SOGA and MOGA based k-means produced a cluster size of 450 each respectively. 449 out of 450 of them belong to the predefined tag of "Great Barrier Reef" and with one irrelevant tag detected respectively. Meanwhile, the generic k-means produced a cluster size of 508. 406 out of 508 of them belong to the predefined tag of "Donald Trump", 60 of them belong to the predefined tag of "Najib Razak" and with 42 irrelevant tags detected. In Table VIII, SOGA based k-means produced a cluster size of 466 respectively. 396 out of 466 of them belong to the predefined tag of "Donald Trump", 49 of them belong to the tag "Najib Razak" and with 21 irrelevant tags. MOGA based k-means produced a cluster size of 500 each respectively. 405 out of 500 of them

TABLE VI: Top 5 representative words extracted from the clusters when k=5

	Topics	Top 5 Words	Relevant Tags	Irrelevant Tags
SOGA k-means	Great Barrier Reef	Reef, Barrier, Great, Australia, Coral	449	1
MOGA k-means	Great Barrier Reef	Reef, Barrier, Great, Australia, Coral	449	1
Generic k-means	Note 7	Note, Phone, Ban, Flight, Take	414	3

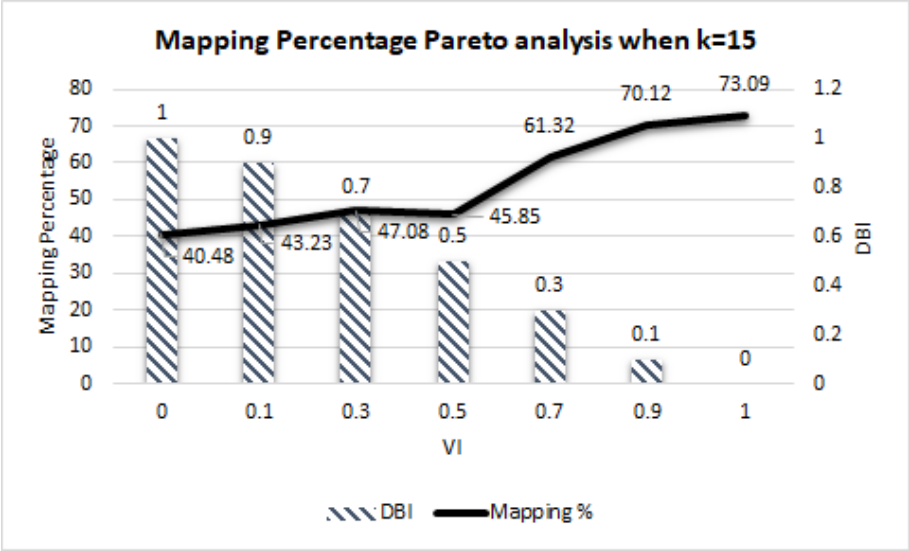

Fig. 6: Mapping Percentage Pareto analysis when k=15 and P_c=0.5

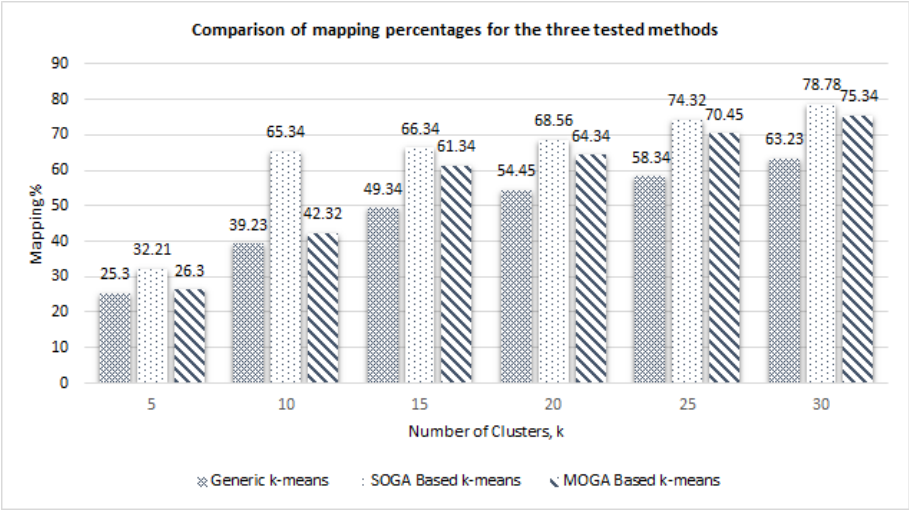

Fig. 7: Comparison of mapping percentages for the three tested methods

TABLE VII: Top 5 representative words extracted from the clusters when k=10

	Topics	Top 5 Words	Relevant Tags	Irrelevant Tags
SOGA k-means	Great Barrier Reef	Reef, Barrier, Great, Australia, Coral	449	1
MOGA k-means	Great Barrier Reef	Reef, Barrier, Great, Australia, Coral	449	1
Generic k-means	Donald Trump, Najib Razak	Trump, Donald, Obama, William, Elect	466	42

TABLE VIII: Top 5 representative words extracted from the clusters when k=15

	Topics	Top 5 Words	Relevant Tags	Irrelevant Tags
SOGA k-means	Donald Trump, Najib Razak	Trump, Donald, Obama, William, Elect	445	21
MOGA k-means	Donald Trump, Najib Razak	Trump, Donald, Obama, Elect, William	459	41
Generic k-means	Donald Trump, Najib Razak	Trump, Donald, Obama, Elect, William	466	42

belong to the predefined tag of "Donald Trump", 54 of them belong to the tag "Najib Razak" and with 41 irrelevant tags. The generic k-means produced a cluster size of 508 each respectively. 406 out of 508 of them belong to the predefined tag of "Donald Trump", 60 of them belong to the tag "Najib Razak" and with 42 irrelevant tags.

V. Conclusion

Based on the results obtained, it is proven that GAs has the ability to provide better solutions when embedded into the k-means clustering solution. The major weakness of k-means clustering algorithm, which is the choice of initial centroids, is addressed via the implementation of a GA based k-means algorithm. The proposed MOGA based k-means clustering

algorithm also makes it feasible to be applied to solve real-world problems which are typically multiobjective in nature. Topics extraction using the proposed algorithm also makes it possible to extract meaningful topics from the ever growing social media sites like Twitter.

For future works, there are still many aspects that can affect the performance of MOGA and the performance of topics extraction. In terms of MOGA, these aspects include using different fitness functions that measure different criteria of the cluster, using a different selection scheme, crossover operator, and mutation operator. Another important aspect is also to consider using the Pareto optimal approach to solve a multiobjective optimization problem, which is believed to have the potential to deliver more promising results. In terms of topics extraction, a different preprocessing approach can be considered to correct any misspelled words instead of removing them. This will greatly improve the quality of the actual content and provide more accurate meaning and insights to the problem. Furthermore, more works can be done in expanding the term in order to allow it to hold more representation of the same meaning.

The main reason to this is that too few words will hinder the performance of the clustering process. Throughout this research, few combinations of parameters are applied and tested to SOGA and MOGA. It is found that the performances of SOGA and MOGA in terms of mapping percentage are able to outperform the generic and un-optimized k-means clustering algorithm. The novelty of the mapping percentage is that it allows maximum number of possible pairings between predefined clusters and produced clusters to be considered and to make sure that both clustering are parallel. The performance of MOGA is slightly below par relative to SOGA. The results also showed when using SOGA and MOGA, the clustering algorithms are capable of producing meaningful clusters and the topics extraction process showed a great number of tweets corresponding to the predefined tags which are relevant to the top five most representative words extracted from each cluster.

REFERENCES

[1] "There are more social media users today than there were people in 1971," https://blog.hootsuite.com/simon-kemp-social-media/, accessed: 2020-03-26.

[2] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14 – 46, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705115002336

[3] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, no. 1, p. 91, Oct 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0254-8

[4] S. L. Johnson, P. Gray, and S. Sarker, "Revisiting is research practice in the era of big data," *Information and Organization*, vol. 29, no. 1, pp. 41 – 56, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S147177271830006X

[5] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on big data in marketing: A text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1 – 7, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2444883417300268

[6] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *Journal of Big Data*, vol. 6, pp. 1–16, 2019.

[7] F. X. Jian, W. Yajiao, and D. Yuan-yuan, "Microblog topic evolution computing based on lda algorithm," *Open Physics*, vol. 16, pp. 509 – 516, 2018.

[8] Y. Jia, S. Li, and R. biao Wu, "Incorporating background checks with sentiment analysis to identify violence risky chinese microblogs," *Future Internet*, vol. 11, p. 200, 2019.

[9] B. Navarro-Colorado, "On poetic topic modeling: Extracting themes and motifs from a corpus of spanish poetry," *Frontiers in Digital Humanities*, vol. 5, p. 15, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdigh.2018.00015

[10] X. Li and K. Wong, "Evolutionary multiobjective clustering and its applications to patient stratification," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1680–1693, 2019.

[11] M. Garza-Fabre, J. Handl, and J. Knowles, "An improved and more scalable evolutionary approach to multiobjective clustering," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 4, pp. 515–535, 2018.

[12] S. Zhu and L. Xu, "Many-objective fuzzy centroids clustering algorithm for categorical data," *Expert Systems with Applications*, vol. 96, pp. 230 – 248, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741730828X

[13] R. Shang, W. Zhang, F. Li, L. Jiao, and R. Stolkin, "Multi-objective artificial immune algorithm for fuzzy clustering based on multiple kernels," *Swarm and Evolutionary Computation*, vol. 50, p. 100485, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2210650218300634

[14] K. K. Bun and M. Ishizuka, "Topic extraction from news archive using tf*pdf algorithm," *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.*, pp. 73–82, 2002.

[15] H. F. Ma and H. L. Ma, "Combining burst detection for hot topic extraction," in *Computational Materials Science*, ser. Advanced Materials Research, vol. 268. Trans Tech Publications Ltd, 8 2011, pp. 1283–1288.

[16] K. Chen, L. Luesukprasert, and S. T. Chou, "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 8, pp. 1016–1025, Aug 2007.

[17] E. Zhou, N. Zhong, and Y. Li, "Hot topic detection in professional blogs," in *Active Media Technology*, N. Zhong, V. Callaghan, A. A. Ghorbani, and B. Hu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 141–152.

[18] S. B. Basri, R. Alfred, and C. K. On, "Automatic spell checker for malay blog," in *2012 IEEE International Conference on Control System, Computing and Engineering*, Nov 2012, pp. 506–510.

[19] R. Alfred, A. Mujat, and J. H. Obit, "A ruled-based part of speech (rpos) tagger for malay text articles," in *Intelligent Information and Database Systems*, A. Selamat, N. T. Nguyen, and H. Haron, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 50–59.

[20] R. Alfred, L. C. Leong, C. K. On, P. Anthony, T. S. Fun, M. N. B. Razali, and M. H. A. Hijazi, "A rule-based named-entity recognition for malay articles," in *Advanced Data Mining and Applications*, H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, and W. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 288–299.

[21] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "A literature review and discussion of malay rule - based affix elimination algorithms," in *The 8th International Conference on Knowledge Management in Organizations*, L. Uden, L. S. Wang, J. M. Corchado Rodríguez, H.-C. Yang, and I.-H. Ting, Eds. Dordrecht: Springer Netherlands, 2014, pp. 285–297.

[22] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 1128–1137. [Online]. Available: https://www.aclweb.org/anthology/P10-1115

[23] M. Okamoto and M. Kikuchi, "Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries," in *Information Retrieval Technology*, G. G. Lee, D. Song, C.-Y. Lin, A. Aizawa, K. Kuriyama, M. Yoshioka, and T. Sakai, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 181–192.

[24] P. Hung, Lai, Rayner, and Alfred, "An optimized multi-layer ensemble framework for sentiment analysis," in *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 2019, pp. 158–163.

[25] J. Li, M. Liao, W. Gao, Y. He, and K.-F. Wong, "Topic extraction from microblog posts using conversation structures," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2114–2123. [Online]. Available: https://www.aclweb.org/anthology/P16-1199

[26] M. K. Islam and M. M. Ahmed, "i-codas: An improved online data stream clustering in arbitrary shaped clusters," *Engineering Letters*, vol. 27, no. 4, pp. 752–762, 11 2019.

[27] S. Ajorlou, I. Shams, and K. Yang, "A fast clustering algorithm for mining social network data," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, vol. 1, London, U.K., July 2014, pp. 106–109.

[28] I. T. Afolabi, O. Y. Sowunmi, and T. Adigun, "Semantic text mining using domain ontology," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, San Francisco, USA, October 2019, pp. 309–314.

[29] T. Semangern, W. Chaisitsak, and T. Senivongse, "Identification of risk of cyberbullying from social network messages," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, San Francisco, USA, October 2019, pp. 276–282.

[30] C. Raposo, C. H. Antunes, and J. P. Barreto, "Automatic clustering using a genetic algorithm with new solution encoding and operators," in *Computational Science and Its Applications – ICCSA 2014*, B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan, and O. Gervasi, Eds. Cham: Springer International Publishing, 2014, pp. 92–103.

[31] T. Vo-Van, T. Nguyen-Thoi, T. Vo-Duy, V. Ho-Huu, and T. Nguyen-Trang, "Modified genetic algorithm-based clustering for probability density functions," *Journal of Statistical Computation and Simulation*, vol. 87, no. 10, pp. 1964–1979, 2017. [Online]. Available: https://doi.org/10.1080/00949655.2017.1300663

[32] R. Alfred, G. J. Chiye, Y. Lim, C. K. On, and J. H. Obit, "A multi-objectives genetic algorithm clustering ensembles based approach to summarize relational data," in *SCDS*, 2016.

[33] R. Alfred, G. Chiye, J. Obit, M. Hijazi, C. On, and H. Lau, "A genetic algorithm based clustering ensemble approach to learning relational databases," *Advanced Science Letters*, vol. 21, pp. 3313–3317, 10 2015.

[34] R. Alfred, "Feature transformation: a genetic-based feature construction method for data summarization." *Computational Intelligence*, vol. 26, pp. 337–357, 08 2010.

[35] Haviluddin and R. Alfred, "A genetic-based backpropagation neural network for forecasting in time-series data," in *2015 International Conference on Science in Information Technology (ICSITech)*, Oct 2015, pp. 158–163.

[36] Y. Liu, X. Wu, and Y. Shen, "Automatic clustering using genetic algorithms," *Applied Mathematics and Computation*, vol. 218, no. 4, pp. 1267 – 1279, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0096300311008204

[37] U. Maulik, S. Bandyopadhyay, and A. Mukhopadhyay, *Multiobjective Genetic Algorithms for Clustering - Applications in Data Mining and Bioinformatics.*, 01 2011.

[38] B. J. D. Sitompul, O. S. Sitompul, and P. Sihombing, "Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm," *Journal of Physics: Conference Series*, vol. 1235, p. 012015, jun 2019. [Online]. Available: https://doi.org/10.1088%2F1742-6596%2F1235%2F1%2F012015

[39] X. Yu, Y. Lu, and X. Yu, "Evaluating multiobjective evolutionary algorithms using mcdm methods," 2018.

[40] S. Wikaisuksakul, "A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering," *Applied Soft Computing*, vol. 24, pp. 679 – 691, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494614004013

[41] R. Alfred, L. C. Leong, and J. H. Obit, "An evolutionary-based term reduction approach to bilingual clustering of malay-english corpora," in *Advances in Information and Communication Technology*, M. Akagi, T.-T. Nguyen, D.-T. Vu, T.-N. Phung, and V.-N. Huynh, Eds. Cham: Springer International Publishing, 2017, pp. 132–141.

[42] M. N. Murty, B. Rashmin, and C. Bhattacharyya, *Clustering Based on Genetic Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 137–159. [Online]. Available: https://doi.org/10.1007/978-3-540-77467-9_7

[43] J. C. Rojas Thomas, M. Mora, and M. Santos, "New version of davies-bouldin index for clustering validation based on hyper rectangles," vol. 2014, 01 2014, pp. 13 (6 .)–13 (6 .).

[44] J. Wu, J. Chen, H. Xiong, and M. Xie, "External validation measures for k-means clustering: A data distribution perspective," *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6050 – 6061, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408004028

[45] L. Fortuna, G. Rizzotto, M. Lavorgna, G. Nunnari, M. G. Xibilia, and R. Caponetto, *Evolutionary Optimization Algorithms*. London: Springer London, 2001, pp. 97–116. [Online]. Available: https://doi.org/10.1007/978-1-4471-0357-8_6

Rayner Alfred is an Associate Professor of Computer Science at the Faculty of Computing and Informatics, Universiti Malaysia Sabah in Malaysia that focuses on Data Science and Software Engineering programmes. He leads and defines projects around knowledge discovery, information retrieval and machine learning that focuses on building smarter mechanism that enables knowledge discovery in structured and unstructured data. His work addresses the challenges related to big data problem: How can we create and apply smarter collaborative knowledge discovery and machine learning technologies that bridge the structured and unstructured data mining and cope with the big data problem. Rayner completed his PhD in 2008 looking at intelligent techniques using machine learning to model and optimize the dynamic and distributed processes of knowledge discovery for structured and unstructured data. He holds a PhD degree in Computer Science from York University (United Kingdom), a Master degree in Computer Science from Western Michigan University, Kalamazoo (USA) and a Computer Science degree from Polytechnic University of Brooklyn, New York (USA) where he was the recipient of the Myron M. Rosenthal Academic Achievement Award for the outstanding academic achievement in Computer Science in 1994. He has authored and co-authored more than 150 journals/book chapters and conference papers, editorials, and served on the program and organizing committees of numerous national and international conferences and workshops.

Rayner is currently a member of IEEE, a Certified Software Tester (CTFL) from the International Software Testing Qualifications Board (ISTQB), and also a certified IBM DB2 Academic Associate (IBM DB2 AA). He leads the Advanced Machine Intelligence (AMI) research group in UMS and he has lead several projects related to knowledge discovery and machine learning on Big Data. Rayner is also the recipient of the Research Fellow at Japan Advanced Institute of Science and Technology (JAIST), Japan. He is also the recipient of multiple GOLD and SILVER awards at national and international research exhibitions in Data Mining and Machine Learning based solutions (Face Recognition and Knowledge Discovery), that include International Trade Fair Ideas in Nuremberg, Germany (iNEA2018) International Invention Innovation Competition in Toronto, Canada (iCAN 2018), Seoul International Invention Exhibition in Seoul, Korea (SIIF 2010). He has secured RM6,931.433.00 worth of project grants.

Loo Yew Jie is a Senior Insight Analyst at Aimia Inc. He received the BSc degree in Computer Science (Hons) from Universiti Malaysia Sabah in 2017 and MSc. degree in Business Intelligence and Analytics from Universiti Teknologi Malaysia in 2020.

Joe Henry Obit is an Associate Professor of Computer Science, department of Data Science at Universiti Malaysia Sabah. His main research interest lies at the interface of Operational Research and Computer Science. In particular, the exploration and development of innovative Operational Research, Artificial Intelligence, and Distributed Artificial Intelligence models and methodologies for automatically producing high quality solutions to a wide range of real world combinatorial optimization and scheduling problems. Dr. Joe Obtained his PhD in Computer Science from the School of Computer Science at the University of Nottingham. His PhD thesis is developing a Novel Meta-heuristic, Hyper-heuristic and Cooperative Search.

Yuto LIM received the B.Eng. (Hons) and M.Inf. Technology degrees from Universiti Malaysia Sarawak (UNIMAS), Malaysia in 1998 and 2000, respectively. He received the Ph.D. degree in communications and computer engineering from Kyoto University in 2005. In October 2005, he was a visiting researcher at Fudan University in China for two months. In November 2005, he was an expert researcher at National Institute of Information and Communications Technology (NICT), Japan until September of 2009. In 2005, he is actively joining the standardization activities of IEEE 802.11s Mesh Networking. He and his team members have introduced two proposals, which currently adopted in the draft of IEEE 802.11s D1.04. He also led the RA-OLSR group in resolving a part of the comments. In 2006, he is also actively joining the standardization activities of next-generation home networks from Telecommunication Technology Committee (TTC), Japan. In 2007, he obtained a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for three years. Since October 2009, he has been working at Japan Advanced Institute of Science and Technology (JAIST) as an associate professor. He is a recipient of IEICE Technical Committee on Energy Engineering in Electronics and Communications Presentation Award for Young Engineers (2009). He is a member of IEEE, IEICE, and IPSJ.

Haviluddin was born in Loa Tebu, East Kalimantan, Indonesia. He graduated from STMIK WCD Samarinda in the field of Management Information, and he completed his Master at Universitas Gadjah Mada, Yogyakarta in the field of Computer Science. He holds a PhD degree in Computer Science from the Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia. He is now a member of the Institute of Electrical and Electronic Engineers (IEEE), International Association of Computer Science and Information Technology (IACSIT), Institute of Advanced Engineering and Science (IAES), Association of Computing and Informatics Institutions Indonesia (APTIKOM) societies.

Azreen Azman received the bachelor's degree in information technology majoring in information systems engineering from Multimedia University, Malaysia, in 1999, the Diploma degree in software engineering from the Institute of Telecommunication and Information Technology, in 1997, and the Ph.D. degree in computer science specializing in information retrieval with the University of Glasgow, Scotland, in 2007. He was accepted directly to second year with Multimedia University. After serving in the industry for a few years, he enrolled for the Ph.D. degree in January 2003. He is currently an Associate Professor with the Universiti Putra Malaysia. His current research interests include information retrieval, text mining, natural language processing, and intelligent systems. He serves as a Committee Member of the Malaysian Society of Information Retrieval and Knowledge Management (PECAMP) and the Malaysian Information Technology Society (MITS).