Concept Trending of Social Media Data Using Apriori Algorithm

Rayner Alfred, *Member, IAENG*, Loh Boon Jing, Joe Henry Obit, Yuto Lim, Haviluddin Haviluddin, Raymond Alfred

***Abstract*—Topic trending is a popular research topic in recent years since, there are massive participations in social web sites, countless number of updates, news, opinions and product reviews are being constantly posted every day. The identification of popular topics discussed or posted on social media platforms is becoming more important as the new knowledge can be extracted from these findings. In this work, a novel method is proposed to extract popular topics from social media and determine the topic trending based on timeline using the Apriori algorithm. The approach uses Twitter's tweets as the dataset. The data is then pre-processed by undergoing several processes that include stop words removal, stemming, tokenization and Term Frequency-Inverse Document Frequency ($TF - IDF$) weighting. The k-means clustering is then performed to cluster each data that consists of processed keywords and collected every day. The popular topics will be then extracted from the clusters and the topic trends will be determined based on the observed frequent patterns and correlation between keywords by using the association rules. The performance of the proposed method is evaluated based on the similarity of the results with the current trends obtained from the Twitter site. The result from the findings shows that the proposed method is able to produce more enriched trends that are similar to current initial trends.**

***Index Terms*—Association Rules, Concept Trending, Clustering, Social Media Mining, Topics Extraction.**

I. Introduction

TOPIC trending has become a significant research in recent years. Since constructing useful trends definitions will contribute towards a better understanding of the interactions in the context of social media, there were many techniques proposed from the researchers and are implemented in real life [1]. Techniques such as topic detection and tracking [2][3][4][5], trend mining-total from partial (TM-TFP) [6], hierarchical clustering algorithm [7], combination of TF*PDF (Term Frequency and Proportional Document Frequency) and Aging Theory [8] are the recent techniques used for trends detection [9][10]. Topic detection and tracking algorithms can be used to cluster, organize and retrieve unstructured information and it is an important research task for information retrieval that can enable users to detect relationships between concepts [11][12]. From these researches, the techniques mentioned in their works can be used to perform popular topics extraction effectively. However, these ideal approaches pay less attention on the frequent patterns based on time series and the correlation of these keywords during a specific period of time that can be produced in order to get more enrich and meaningful trends.

The aim of this study is to propose a method that can be used to extract popular topics from social media and determine the trends based on the timeline by coupling k-means clustering and association rules mining. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. The frequent pattern mining leads to the discovery of association and correla-tions among items in large transactional or relational data sets.

The contribution of this paper may thus be summarized as: (i) to design and propose a method to capture social media opinions and perform pre-processing processes on these social media opinions, (ii) to design and propose a clustering and extraction method in order to extract popular topics from these social media opinions based on timeline and (iii) to design and assess the association rules algorithm that is used to produce trending by extracting popular topics based on timeline.

The rest of this paper is organized as follows. Some related works are described in Section 2. The proposed approach is described in Section 3. Section 4 presents the result and discussion on the proposed technique. Section 5 concludes this paper by summarizing the findings and future research directions that can be undertaken.

II. Related Works

Trends in social networks have recently been a major focus of interest among researchers [13][14]. Social networks provide large-scale information infrastructures for people to discuss and exchange ideas about different topics. This will consequently generate all kinds of information. Since there are massive information produced every day, appropriate data mining techniques is required to analyze such large, complex, and frequently changing social media data. Studying these trends may discover the emergent or suspicious behavior in the network. They can also be viewed as a refletion of societal concerns or even as a consensus of collective decision making. This will lead to the understanding of how the community "thinks" [15].

Manuscript received October 7, 2021; revised October 28, 2022.

This work was conducted by the Advanced Machine Intelligence research group and supported by Universiti Malaysia Sabah.

R. Alfred is an associate professor of Computer Science at the Advanced Machine Intelligence Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah (e-mail: ralfred@ums.edu.my).

B.J. Loh is an undergraduate student at the Faculty of Computing and Informatics, Universiti Malaysia Sabah (e-mail: bonbonloh93@hotmail.com).

J.H. Obit is an associate professor of Computer Science at the Advanced Machine Intelligence Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah (e-mail: joehenry@ums.edu.my).

Y. Lim is an associate professor of Computer Science at the 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 an associate professor of Computer Science at the Department of Informatics, Universitas Mulawarman, Samarinda, Indonesia (e-mail: haviluddin@unmul.ac.id).

A. Raymond is a research consultant at the Allyssa Certification Sdn. Bhd., 88450 Kota Kinabalu, Sabah, Malaysia (e-mail: raymond_alfred@yahoo.my)

In future, the main challenge is to handle the massive amount of information obtained from the social media. Twitter is one of the evolving social media which, on average, hosts around 200 million tweets per day [16][17]. This includes products, ser-vices, social issues, news, incidents and reviews etc. Data generated from social network such as Twitter and Facebook is vast, noisy, distributed and dynamic [18][19]. Since discovering the trends may help in various sectors, these vast and noisy information must be filtered for any meaningful utilization. Currently, the methods used for topic extraction so far are inefficient based on non-correlated keywords detection and machine learning techniques [20][21][22]. An efficient technique for topics extraction and detecting multiple trends based on timeline will become one of the challenges for the researchers to extract the main opinion from these large text documents.

The algorithms proposed and applied in topics extraction and trends detection were able to detect hot topics and track them to a good accuracy [23][24]. While, a Total-From-Partial (TFP) algorithm and Self Organizing Maps (SOM) have been proposed in which the proposed algorithm is able to define interesting trend [25]. Chen's research encourages participants to play roles as tag creators for contributing key-words for taxonomy, collaborative, and social purposes, and as tag consumers for knowing what other people are interested in and how they characterize the same re-sources [26][27]. In addition to that, Nguyen has introduced the Combination of TF-PDF and Aging Theory that is able to identify the popular information about which tech-nologies were invented with high intention from a huge number of patents with lengthy and difficulty technical terms [28]. In Aiello's research, six topic detection methods were compared and the factors of influencing the performance of the meth-ods were identified [29]. Although these existing techniques are shown to have well-performing results on the selected dataset, there are still some deficiencies on each of the research. Based on the review of previously published works, most algorithms deal with single keyword trending, not many of the researched conducted focus on multiple correlated keywords trending. In the proposed framework of this pa-per, frequent patterns and association rule mining will be implemented in order to improve the trend detection from the twitter data. The evaluation of the implementation will be made.

Many researches have been conducted to discover the techniques to efficient topic extraction and trend detection. Association rules mining [30] is one of the interesting method that cab used for multiple trends association de-tection. Based on the frequent patterns produced by the topics extracted, the rules will show the correlation between the topics and produce a more enriched trends. Therefore, this paper propose a new method in which multiple topics trending can be extracted by finding the frequent patterns using Apriori association rules and compute the correlation between keywords to produce multiple topics extraction and trending based on correlated keywords. The full framework of the proposed algorithm is shown in Fig. 1

III. Methodology

This aim of this work is to implement and evaluate a new method in which multiple topics trending can be extracted and retrieved by finding frequent patterns across the timeline and compute the correlation between keywords to produce multiple topics trending based on correlated keywords. Basically, there are five steps in the proposed framework, which are data collection, data pre-processing, text clustering, popular topics extraction and determine topic trending as shown in Fig. 1.

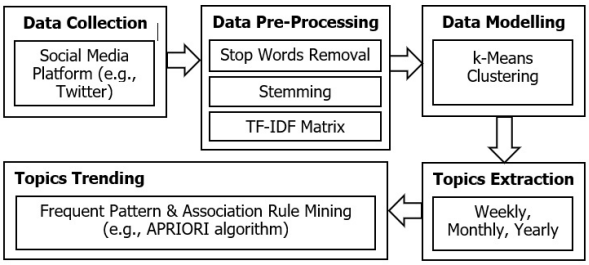

Fig. 1: The full framework of the proposed concept trending using association rules

A. Data Collection

Social media opinions are crawled and collected from Twitter site. Twitter data is interesting because the tweets occur at the "speed of thought" and are available for consumption in real time. Due to its openness for public consumption, the data can be obtained from anywhere in the world. There are three reasons that Twitter is selected: a) Twitter's data is already in a convenient format for analysis. b) Twitter's API is well designed and easy to access. c) Twitter's terms of use for the data are relatively liberal as compared to other APIs. The twitter data were collected every day for 4 weeks. One set of text data comprises of twitter texts that have been posted for one day only. After four weeks, there will be 7 sets of text data collected per week.

The data collection was parfomed based on the timeline. The period of data collection was made between 20th March 2017 and 17th April 2017 (a total of 28 days). There were 1000 tweets collected from the public stream per day (a total of 28,000 tweets for 28 days). Once the data were collected, the tweets texts are saved as a comma separated values (CSV) file. The collected data was then processed using several text pre-processes (e.g., named-entity recognition, part-of-speech tagging, stemming and term reduction [31][32][33][34]) and stored as text-document matrix based on the minimum threshold value of TF-IDF (Term Frequency – Inverse Doc-ument Frequency)[35]. All these processes will be discussed in the next section.

B. Data Pre-processing

The collected data will then be transformed into a repre-sentation suitable for applying the data modelling through data pre-processing. The tweet texts are then cleaned and standardized using upper case letter conversion to lower case letter, white spaces and non-alphabetical characters removal [36]. Next, the text document will be tokenized. Tokenization is a process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. Tokens are separated by whitespace, punctuation

marks or line breaks. After the tokenization process, several text pre-processes will be performed such as the named-entity recognition, part-of-speech tagging and stop words removal [37], stemming for term reduction. Fig 2 illustrates all the pre-processing tasks performed in order to prepare the documents for clustering and concept trending extraction using the Apriori algorithm.

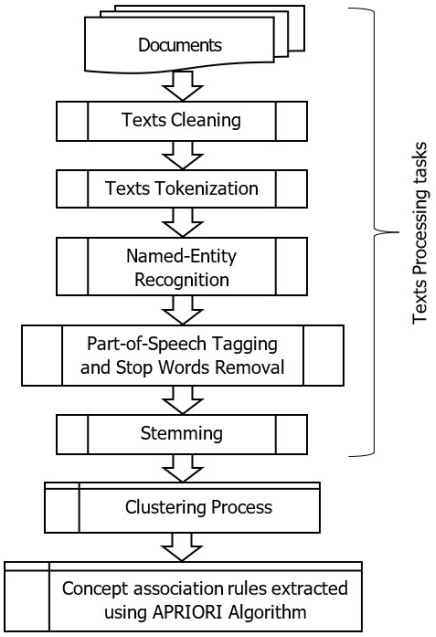

Fig. 2: Several preprocessing tasks used to prepare documents for clustering

The *Named-Entity Recognition* (NER) process is a process that seeks to locate and classify named entities into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. In this work, we apply the Conditional random fields (CRF) to perform the NER process. The CRF is a statistical sequence modeling framework first introduced in [38]. CRFs are normally used for the prediction and analysis of labels for data in natural language writing. In this model, $X = \{x_1, x_2, x_3, ..., x_T\}$ are the input data in which components are connected in sequence, and $Y = \{y_1, y_2, y_3, ..., y_T\}$ are the labels for each component of the input data. In other words, when a new x is given, a y value is predicted using the following model:

$$p(y|x) = \frac{1}{z(x)} \prod_{t=1}^{T} exp\ \{\sum_{k=1}^{k} w_k f_k(y_t, y_{y-1}, x_t)\} \quad (1)$$

$$z(x) = \sum_{y} exp\ (\sum_{k} w_k f_k(y, x)) \quad (2)$$

where $z(x)$ standardizes the probability value, and f_k is a feature function, which is a characteristic function on feature k. This function returns 1 when the given input y_t, y_{t-1}, x_t includes a feature k, and returns 0 otherwise. w_k is the weight of the feature.

The *Part of Speech tagging* process (POS) is used to mark up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context. In other words, POS is the process of identifying words as nouns, verbs, adjectives, adverbs, etc. Based on the results obtained, all the stop words will be removed in order to reduce the dimensionality of the document. Stop Words are words which do not carry important significance to be used in information retrieval. These stop words include ”a”, ”the”, ”actually”, accordingly, etc.

The *Stemming* process is used to reduce inflected (or sometimes derived) words to their word stem, base or root form-generally a written word form. For instance, all the words ”player”, ”played”, ”playing”, and ”plays” can be reduced to its root word ”plays”. The most common stemmer algorithm is called Porter stemmer algorithm which is an essential tool for natural language processing in the area of information access [39]. There are six sub-phases of the Porter stemmer algorithm as follow:

1) Get rid of plurals and $-ed$ or ing.
2) Turns terminal y to i when there is another vowel in the stem.
3) Maps double suffixes to single ones, ization, ational.
4) Deal with suffixes, $-full$, $-ness$, etc.
5) Takes off $-ant$, $-emce$, etc.
6) Removes a final $-e$.

The remaining texts or tokens will then be used in the process which is the construction of term-document matrix in which each token will be weighted using the Term Frequency-Inverse Document Frequency (TF-IDF) as a term weighting scheme. TF-IDF is a numerical statistic which reveals how important a word is to a document in a collection. The value of TF-IDF increases proportionally to the number of times a word appears in the document, but inversely proportional to the frequency of the word in the corpus. TF-IDF is the product of two statistics which are term frequency and inverse document frequency. Term Frequency (TF) is defined as the number of times a term appears in a document [40][35].

$$tf(t,d) = 1 + \log_2 f_{t,d}, f_{t,d} > 0 \quad (3)$$

$$tf(t,d) = 0, f_{t,d} = 0 \quad (4)$$

where $tf(t,d)$ is the number of occurrences of term t in document d. Inverse Document Frequency (idf) is used to measure the importance of a term in a text document collection. The idf is then defined as

$$idf(t) = \log_2 \frac{N}{n_t} \quad (5)$$

where N is the total number of documents in the corpus and n_t is the number of documents that contain at least one occurrence of term. The tf_idf is calculated as

$$tf_idf(t,d) = (1 + \log_2 f_{t,d}) \times \log_2 \frac{N}{n_t} \quad (6)$$

Once, the term-document matrix has been constructed, then we can use this data representation for the data modelling, which is the documents or texts clustering.

C. Data Modelling: Texts Clustering

In the proposed framework, k-means Clustering algorithm is applied to summarize text data into few clusters for each day. Keywords are identified in each cluster. In general, k-means algorithm is known as the simplest unsupervised

learning algorithms to solve clustering problem. The idea of k-means algorithm is that k clusters can be represented by the mean of the document assigned to that cluster, which is called the centroid of that cluster [41]. k-means clustering is a top-down algorithm that clusters the objects into k number of groups with regard to its attributes or features, where k is a positive integer number and specified by users [42]. The basic steps of clustering process include;

1) Select k points at random as cluster centers.
2) Assign objects to their closest cluster center according to a specific distance method (e.g., Euclidean distance function).
3) Calculate the centroid or mean of all objects in each cluster.
4) Go to Step 2 if the membership of each object has changed, otherwise, stop.

One of the reasons for the popularity of k-means algorithm is because of the linear complexity. This means that if the number of instances is substantially large, this algorithm is computationally attractive. k-means algorithm is also famous for its ease of interpretation, simplicity of implementation, speed of convergence and adaptability to sparse data. However, due to its sensitivity to noisy data and outliers, the squared error may have significant increase even with a single outlier. It is applicable only when the mean is defined and the number of clusters is required in advance. The k-means clustering is implemented to cluster the terms extracted. In order to have the optimized number of clusters, k, the Elbow method is used. Table I shows the optimum k number estimated for the 28 days. A genetic algorithm based clustering can also be used to optimize the number of clusters [43][44][45]. For instance, for day 1 in week 1, the optimum number of cluster, k, is 100.

TABLE I: The optimum k number of clusters used for 28 days

Day	Week 1	Week 2	Week 3	Week 4
1	100	75	100	99
2	81	67	106	94
3	76	85	95	113
4	79	64	117	115
5	74	76	93	79
6	70	102	97	70
7	78	102	101	101
Total Clusters	558	571	708	671

D. Extracting Topics from Clusters

After clustering is performed, a few clusters of keywords are produced. These clusters are illustrated by using the cluster plot shown in Fig. 3. The graph shown in Fig. 3 illustates the 6 clusters plotted against the first two principal components of the data and which cluster each object belongs to. These two components explain 71.36% of the point variability. Since our data are multivariate, thus it is difficult to inspect all the many bivariate scatterplots. For that reason, the scatterplot of the first two principal components were derived from the data. The "71.36% of variability" indicates that more than half of the information about the multivariate data is captured by this plot of components 1 and 2. The 3rd component could be added and this will increase the variability.

Each cluster contains keywords that have high frequencies of occurrence in the documents. These keywords are extracted from each cluster produced earlier. Since there are 7 sets of texts data collected per week, there will be seven sets of clusters results produced and as a result, there will be seven sets of keywords obtained from the clusters produced. These keywords will determine the popular topics among the social media opinions for each week.

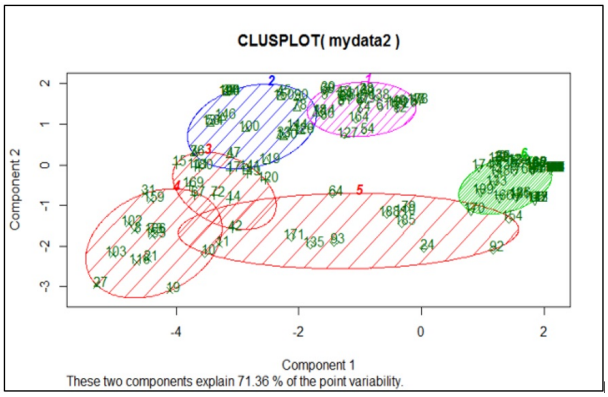

Fig. 3: Example of clusters plot

E. Multiple Topics Trending

A list of frequent keywords extracted over time is generated based on the popular topics extracted. The frequent keywords obtained from each cluster represent hot topics and will be used to determine the topic trending. The topic trends are determined based on the frequent keywords over the time and the correlation between the keywords by using association rules mining. A frequent pattern is a pattern that occurs frequently in the dataset. Pattern can be defined as a set of items, subsequences, subgraphs [46]. In this research, the patterns generated are studied and applied in the association rule mining. Apriori is the algorithm used to extract the frequent pattern of keywords. Based on the Apriori pruning principle, the pattern which is infrequent, its superset should not be tested, which means the pattern will be eliminated.

Then, rules are generated based on these frequent patterns and the correlation between the keywords is identified based on the support and confidence measures. In the end, the topic trend can be detected. Apriori is by far the most well-known method in association rule algorithm for generating frequent pattern. This algorithm uses the property that there must be a large itemset for any subset of a large itemset. It is also assumed that items within an itemset are kept in the order of item's names. In order to avoid too many small candidate itemsets, Apriori generates candidate itemsets by joining the large item-sets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database [47]. The pseudocode for generating candidates is shown as follow:

INPUT: S where S = dataset
OUTPUT: Set of Candidate Itemsets
Require: $S = \emptyset$

```
1. Procedure GenerateCandidates
2.   i ← 2
3.   num ← NumAttributes(S)
4.   candidates [] ← null
4.   support [] ← null
5.   while i < num do
6.     candidates [] ← all sets of size i
6.     support [] ← support(candidates [])
7.     i ← i + 1
8.   end while
9.   end procedure
```

Given a dataset S, we start generating the set of candidate itemsets of size 2, i = 2, and compute the support for these itemsets. Next, we increase the size by 1 and now we are generating the candidate itemsets of size 3, i = 3, and compute the support for these itemsets. This process continues until i equals to the number of attributes or features or terms in the database. One may reduce the number of candidates by filtering only those candidates having equal or greater tham the minimum predefined threshold value for the support measures for all the candidates generated.

Association rule mining is one of the data mining applications that has been proven to be quite useful in the marketing and retail communities as well as other more diverse fields. It is used to identify interesting correlation relationships among a set of items in a database [30]. Rule support and confidence are two measures of rule interestingness. Basically, association rules are considered interesting if both of the minimum support threshold and confidence threshold are satisfied, while the threshold can be determined by users or domain experts.

IV. Results and Discussion

A. Multiple Topics Extraction

The cluster for the first-day data is illustrated using the clusplot function provided in the library cluster in R is used. The graph shown in Fig. 4 is the clusplot result of k-means clustering based on the membership of the documents in each cluster.

The clustering algorithms work in a mathematical space whose dimensionality equals the number of words in the corpus. Clearly, this is impossible to visualize. In order to handle this situation, the Principal Component Analysis (PCA) is used to reduce the number of dimensions to 2 (in this case) in such a way that the reduced dimensions capture as much of the variability between the clusters as possible (and hence the comment, "these two components explain 6.5% of the point variability" at the bottom of the plot in Fig. 4)

Since there are more than two terms extracted, Principal Component Analysis (PCA) is performed for the clusplot. In order to increase the point of variability, component 3 is added to the plot and is demonstrated using the 3D plot provided by the function $plot3d()$ from library rgl in R. Fig. 5 shows the 3D plot of the clusters from the front view for the first-day data. While Fig. 6 and Fig. 7 show the top view and the bottom view. Each cluster contains keywords that are related to each document in the cluster. The keywords are extracted from the clusters based on the value in the $tweets\_cluster\$centers$ command.

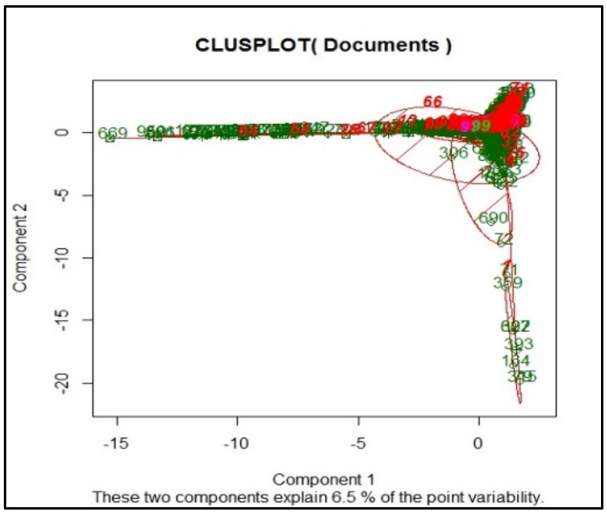

Fig. 4: Clusplot of the clusters based on the membership of documents

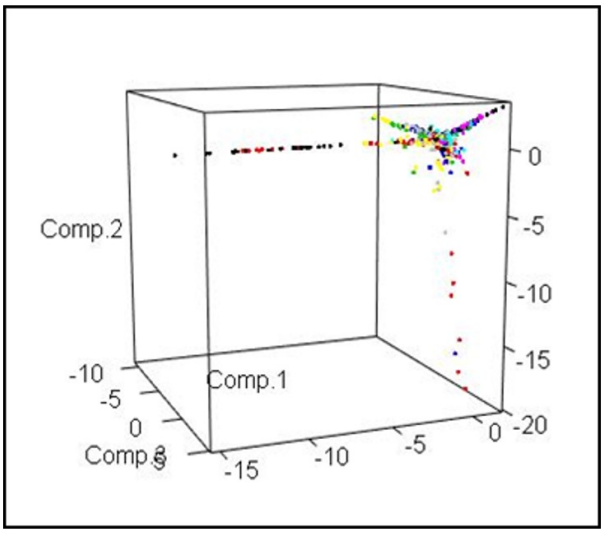

Fig. 5: 3D plot of the clusters based on the membership of documents (front view)

B. Multiple Topics Trending

Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. The association rule is an implication expression of the form $X \rightarrow Y$, where X and Y are itemsets. For instance, this association rule $\{Milk, Diaper\} \rightarrow \{Beer\}$ can be generated from the market basket analysis. The rule evaluation metrics include *Support*, *Confidence* and *Lift*. The *Support* evaluation metric is computed based on the fraction of records that contain both X and Y and the *Confidence* evaluation metric measures how often items in Y appear in records, R, that contain X. The *Support* of item X, $supp(X)$, and the *Confidence* of the rule $X \rightarrow Y$, $conf(X \rightarrow Y)$, evaluation metrics are expressed as:

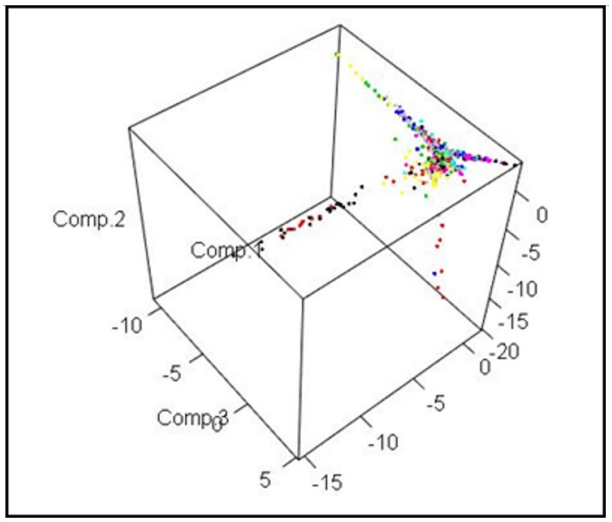

Fig. 6: 3D plot of the clusters based on the membership of documents (top view)

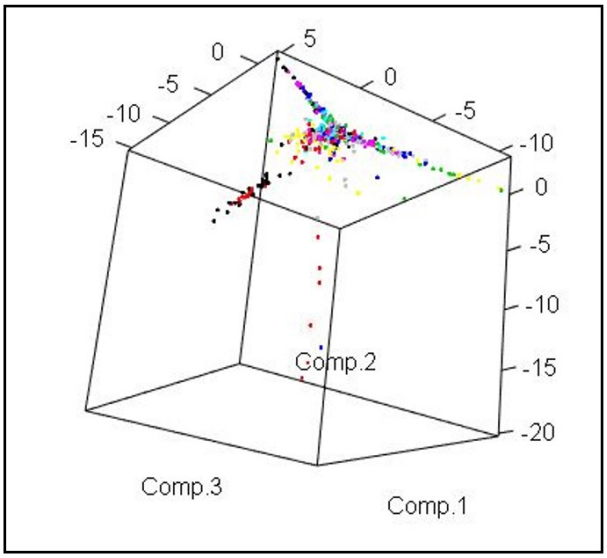

Fig. 7: 3D plot of the clusters based on the membership of documents (bottom view)

$$supp(X) = \frac{\{X \in R\}}{R} \tag{7}$$

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{8}$$

The *Lift* of the rule $X \rightarrow Y$, $Lift(X \rightarrow Y)$, is defined as:

$$Lift(X \rightarrow Y) = \frac{supp(X \rightarrow Y)}{supp(X) \times supp(Y)} \tag{9}$$

In this work, it is considered that if certain topics exist as hot topics, there are also other topics which are more (or less) likely to occur at the same time. For instance, this association rule $\{PrimeMinister, Malaysia\} \rightarrow \{Holidays\}$ can be generated from the news. As a set of frequent keywords is extracted from each cluster, the group of keywords is considered as a group of terms and the number of clusters is considered the number of transactions that exist in the transaction market basket analysis. Then, the association rules can be generated by applying the apriori algorithm which is available in R.

In this research, the lift value is used to find the most important rules generated. Apart from the lift value, the minimum support value and the minimum confidence value are also considered in order to generate the number of rules that need to be generated. In order to avoid too many rules (to avoid less important rules generated) or too little rules (to avoid less information retrieved), the suitable number of rules generated should be adjusted by adjusting the minimum support value and minimum confidence value. Table II shows suitable minimum support value, minimum confidence value and the number of rules generated for each dataset.

TABLE II: Minimum support and confidence values and the number of rules generated for each week dataset

	Week 1	Week 2	Week 3	Week 4
Min Support	0.008	0.008	0.012	0.008
Min Confidence	0.25	0.25	0.25	0.25
Rules generated	2435	1920	2428	2408


```
[
 {
  "url": "http://twitter.com/search?q=%23KimJongNam",
  "query": "%23KimJongNam",
  "tweet_volume": null,
  "name": "#KimJongNam",
  "promoted_content": null
 },
 {
  "url": "http://twitter.com/search?q=Russia",
  "query": "Russia",
  "tweet_volume": 560203,
  "name": "Russia",
  "promoted_content": null
 },
 {
  "url": "http://twitter.com/search?q=%22Ghost+in+the+Shell%22",
  "query": "%22Ghost+in+the+Shell%22",
  "tweet_volume": 44637,
  "name": "Ghost in the Shell",
  "promoted_content": null
 },
 {
  "url": "http://twitter.com/search?q=%23WrestleMania",
  "query": "%23WrestleMania",
  "tweet_volume": 763360,
  "name": "#WrestleMania",
  "promoted_content": null
 },
 {
  "url": "http://twitter.com/search?q=%22SEA+Games%22",
  "query": "%22SEA+Games%22",
  "tweet_volume": null,
  "name": "SEA Games",
  "promoted_content": null
 },
```


Fig. 8: Top five hot topics extracted from the Twitter public stream in Malaysia

The generated association rules related to the trends will be compared to the actual trends collected from the Twitter. The actual trends collected are based on the location selected, which is Malaysia. Fig. 8 shows the top five hot topics extracted from the Twitter public stream in Malaysia from 27th March 2017 to 3rd April 2017. These five hot topics include *Kim Jong Nam*, *Russia*, *Ghost in the Shell*, *WrestleMania* and *Sea Games*. Since there are many association rules generated from the proposed approach, only the top five rules with the highest lift and related to the current trend will be chosen to be compared to the current trends.

Based on Table III, the topics can be generated based on the keywords extracted from the clusters. For the Russia topic, the related keywords are “news”, “nunes”, “trump”, “trial” and “russia”. These keywords indicate the news about Devin Nunes, the chairman of the House of Representatives Permanent Select Committee on Intelligence and an ally of

TABLE III: Association rules generated by the proposed framework

Rule No	Russia
R1	news, nunes, trump $\Longrightarrow$ trial
R2	news, nunes, russia, trump $\Longrightarrow$ trial
R3	news, nunes $\Longrightarrow$ trial
R4	news, nunes, russia $\Longrightarrow$ trial
R5	nunes, russia, trump $\Longrightarrow$ trial
R6	nunes $\Longrightarrow$ russia, trump, trial
R7	nunes, trump $\Longrightarrow$ trial
R8	nunes, russia $\Longrightarrow$ trial
	Ghost in the Shell
G1	ghost in the shell, shell $\Longrightarrow$ ghost
G2	ghost, ghost in the shell $\Longrightarrow$ shell
G3	scarlett, shell $\Longrightarrow$ ghost
G4	ghost, scarlett $\Longrightarrow$ shell
G5	seeing, shell $\Longrightarrow$ ghost
G6	seeing $\Longrightarrow$ shell, ghost
G7	seeing, scarlett, shell $\Longrightarrow$ ghost
G8	seeing, shell $\Longrightarrow$ scarlett, ghost
	Wrestlemania
W1	cena, nikki, tonight $\Longrightarrow$ bella
W2	cena, nikki, tonight $\Longrightarrow$ bella
W3	cena, win $\Longrightarrow$ comeback
W4	cena, ring $\Longrightarrow$ bella
W5	nikki, tonight $\Longrightarrow$ bella
W6	nikki, ring, tonight $\Longrightarrow$ bella
W7	nikki, tonight $\Longrightarrow$ ring, bella
W8	cena, nikki $\Longrightarrow$ bella
	Kim Jong Nam
K1	kim $\Longrightarrow$ north
K2	kim $\Longrightarrow$ korea
K3	kim, north $\Longrightarrow$ korea
K4	kim, korea $\Longrightarrow$ north
K5	north, korea $\Longrightarrow$ kim
K6	kim, north, korea $\Longrightarrow$ war
K7	kim, korea $\Longrightarrow$ north, war
K8	north, korea $\Longrightarrow$ war

TABLE IV: Support, Confidence and Lift values for all rules

Rule No	Support	Confidence	Lift
	Russia		
R1	0.0070	0.417	3.758
R2	0.0092	0.265	3.483
R3	0.0102	0.323	1.596
R4	0.0116	0.354	1.388
R5	0.0124	0.332	1.378
R6	0.0118	0.328	1.375
R7	0.0109	0.367	1.370
R8	0.0123	0.345	1.320
Average	**0.0107**	**0.3414**	**1.9585**
	Ghost in the Shell		
G1	0.0090	0.272	**3.811**
G2	0.0079	0.402	**3.688**
G3	0.0079	0.402	**3.688**
G4	0.0094	0.263	**3.493**
G5	0.0091	0.276	**3.490**
G6	0.0089	0.254	**3.488**
G7	0.0099	0.223	**3.485**
G8	0.0094	0.212	**3.450**
Average	**0.0089**	**0.2880**	**3.5741**
	Wrestlemania		
W1	0.0071	0.431	3.942
W2	0.0079	0.402	3.688
W3	0.0094	0.263	3.493
W4	0.0105	0.318	2.281
W5	0.0102	0.288	1.265
W6	0.0109	0.250	1.264
W7	0.0095	0.274	1.263
W8	0.0102	0.345	1.233
Average	**0.0095**	**0.3214**	**2.3036**
	Kim Jong Nam		
K1	0.0090	0.272	3.801
K2	0.0105	0.317	2.283
K3	0.0105	0.317	2.283
K4	0.0102	0.308	1.596
K5	0.0098	0.305	1.590
K6	0.0110	0.310	1.489
K7	0.0116	0.321	1.475
K8	0.0106	0.311	1.455
Average	**0.0104**	**0.3076**	**1.9965**

Republican President Donald Trump, characterized charges that he made unauthorized disclosure of classified as "entirely false and politically motivated".

While for the Ghost in the Shell topic, this movie's release date is on 30th March 2017 in Malaysia. It is a science fiction action film directed by Rupert Sanders and written by Jamie Moss, William Wheeler and Ehren Kruger, based on the Japanese manga of the sane name by Masamune Shirow. It is about Major (Scarlett Johansson) is the first of her kind: a human saved from a terrible crash, who is cyber-enhanced to be a perfect soldier devoted to stopping the world's most dangerous criminals. When terrorism reaches a new level that includes the ability to hack into people's minds and control them, Major is uniquely qualified to stop it. As she prepares to face a new enemy, Major discovers that she has been lied to: her life was not saved, it was stolen. She will stop at nothing to recover her past, find out who did this to her and stop them before they do it to others. The related keywords are "ghostintheshell", "shell", "ghost", "scarlette" and "animation".

For the KimJongNam topic, it is related to the eldest son of Kim Jong-il, deceased former leader of North Korea. The related keywords are "kim", "north" and "korea". Kim Jong-nam was exiled from North Korea in 2003. He was found dead on 13th February 2017 in Malaysia as the result of a suspected chemical attack. On 30th March 2017, Kim Jong-nam's body had been approved to be released to North Korea by Prime Minister Najib Razak. There are only four rules generated as only a few of tweets related to Kim Jong-nam was retrieved. So, this causes a fewer amount of rules generated to the topic.

Another topic is about the Wrestlemania 2017, the thirty-third annual Wrestle-Mania professional wrestling pay-per-view event produced by WWE. The related keywords are "cena", "nikki", "bella", "ring", "miz", "maryse", "win" and "wwe". John Cena and Nikki Bella beat The Miz and Maryse in their mixed tag team match in Wrestlemnia 33. In the following match, John Cena proposed to Nikki Bella.

For the SEA Games topic, it is related to the 2017 Southeast Asian Games, which officially known as the 29th Southeast Asian Games. It was about the four stadiums involved in the project would be completed in time for the c from 19th August 2017 to 31st August 2017. Due to the reason of fewer amounts of related tweets retrieved, the keywords become unimportant as the TF-IDF values are low. So, the keywords may be eliminated before clustering is performed. Thus, we cannot get the association rules related to SEA Games.

Based on Table IV, the top 8 rules generated for the topic entitled "Ghost in the Shell" have the highest average *Lift* value compared to other rules generated in the other topics. On the other hand, the top 8 rules generated from the topic entitled "Russia" have the lowest average *Lift* value compared to other rules generated in the other topics. It shows that in each topic, rules that are generated could have different maximum value for the rules generated. The results also show that the proposed approach is able to produce the topics trend that are correlated with each other and it was found that these topics are almost similar to the current trends extracted from Twitter. In other words, the proposed technique is not only able to produce the expected topic trends, it also provides more enriched information about the hot topic generated by having these rules.

TABLE V: t-test results (p-value)

	R	G	W	K
R	NA	0.00061	0.55086	0.93579
G	0.00061	NA	0.01098	0.00009
W	0.55086	0.01098	NA	0.56202
K	0.93579	0.00009	0.56202	NA

A t-test is then used to determine whether there is a significant quality (Lift) difference between the sets of rules extracted from the clusters. The null hypothesis (H_0) states that there is no significant quality difference between the sets of rules extracted from the clusters. The alternative hypothesis (H_a) states that there is significant quality difference between the sets of rules extracted from the clusters. Table V shows the t-test results (p-value). There is significant quality difference between the set of rules in set R and G, since the p-value < 0.05. This means that the quality of rules in set G is better than the quality of rules in set R because the rules in set G has higher average Lift value. Similary, The p-value < 0.05 for the t-test between rules in set G and W, and also the t-test between rules in set G and K. We can conclude that the set of rules in set G has the highest quality compared to the other rules in sets R, W and K.

V. Conclusion

Extracting the most discussed topics from the social media platform may enable us to discover some temporal trends for a specific period of time. Trends in social data can be discovered to unveil certain trends related to hot issues or topics that are inter-related to each other. Trends are not only reflecting real-world events, but also drive offline behavior. Many ecommerce businesses utilizes these data to find out what is holding consumer interest and making profit for their business. That is the reason why topic trending becomes a significant research for people to uncover and understand the trends in order to meet a business or organization's specific objectives. The results obtained in this work showed that the proposed technique is able to automatically generate enriched topics trend by using the association rules mining. The keywords that form the rules are able to provide more information related to the hot topics extracted from the Twitter site. The rules indicate the correlation between the keywords and this can produce a topic trend that is related to the current trends. Using this proposed technique, the trends produced are able to provide more information related to the topics and users are able to know more information instead of knowing the title of the topics only.

References

[1] 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.

[2] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, "Research on topic detection and tracking for online news texts," *IEEE Access*, vol. 7, pp. 58 407–58 418, 2019.

[3] A. Rafea and N. A. GabAllah, "Topic detection approaches in identifying topics and events from arabic corpora," *Procedia Computer Science*, vol. 142, pp. 270–277, 2018, arabic Computational Linguistics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918321987

[4] 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

[5] 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.

[6] P. Nohuddin, F. Coenen, R. Christley, and W. Sunayama, "Identification and visualisation of pattern migrations in big network data," vol. 7458, 09 2012, pp. 883–886.

[7] V. Gerla, M. Murgas, A. Mladek, E. Saifutdinova, M. Macas, and L. Lhotska, "Hybrid hierarchical clustering algorithm used for large datasets: A pilot study on long-term sleep data," in *Precision Medicine Powered by pHealth and Connected Health*, N. Maglaveras, I. Chouvarda, and P. de Carvalho, Eds. Singapore: Springer Singapore, 2018, pp. 3–7.

[8] K. Nguyen, Byung-Joo Shin, and Seong Joon Yoo, "Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, Jan 2016, pp. 223–230.

[9] W. Melek, Z. Lu, A. Kapps, and W. Fraser, "Comparison of trend detection algorithms in the analysis of physiological time-series data," *IEEE transactions on bio-medical engineering*, vol. 52, pp. 639–51, 05 2005.

[10] M. Gorawski, A. Gorawska, and K. Pasterak, "The tube algorithm: Discovering trends in time series for the early detection of fuel leaks from underground storage tanks," *Expert Systems with Applications*, vol. 90, pp. 356 – 373, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417305547

[11] 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.

[12] 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.

[13] K. Dong Sung and J. Kim, "Public opinion sensing and trend analysis on social media: A study on nuclear power on twitter," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, pp. 373–384, 11 2014.

[14] M. M. Müller and M. Salathé, "Crowdbreaks: Tracking health trends using public social media data and crowdsourcing," *Frontiers in Public Health*, vol. 7, p. 81, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpubh.2019.00081

[15] C. Budak, D. Agrawal, and A. Abbadi, "Structural trend analysis for online social networks." *PVLDB*, vol. 4, pp. 646–656, 07 2011.

[16] J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," *Information Fusion*, vol. 52, pp. 344 – 356, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253518308704

[17] A. Mehmood, A. Palli, and M. Khan, "A study of sentiment and trend analysis techniques for social media content," *International Journal of Modern Education and Computer Science*, vol. 6, pp. 47–54, 12 2014.

[18] M. Fire and R. Puzis, "Organization mining using online social networks," *Networks and Spatial Economics*, vol. 16, no. 2, pp. 545–578, Jun 2016. [Online]. Available: https://doi.org/10.1007/s11067-015-9288-4

[19] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. Oconnor, M. Paul, and G. Gonzalez, "Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019," 01 2019, pp. 21–30.

[20] D. Milioris, "Trend detection and information propagation in dynamic social networks," 05 2015.

[21] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian Informatics Journal*, vol. 17, no. 2, pp. 199 – 216, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1110866515000651

[22] M. Cordeiro, R. P. Sarmento, P. Brazdil, M. Kimura, and J. Gama, "Identifying, ranking and tracking community leaders in evolving social networks," in *Complex Networks and Their Applications VIII*, H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Eds. Cham: Springer International Publishing, 2020, pp. 198–210.

[23] R. Kanagasabai and A.-H. Tan, "Topic detection, tracking and trend analysis using self-organizing neural networks," vol. 2035, 04 2001, pp. 102–107.

[24] Indra, E. Winarko, and R. Pulungan, "Trending topics detection of indonesian tweets using bn-grams and doc-p," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 2, pp. 266 – 274, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S131915781730280X

[25] N. Khanh Ly and S.-H. Myaeng, "Query enhancement for patent prior-art-search based on keyterm dependency relations and semantic tags," 07 2012, pp. 28–42.

[26] S.-Y. Chen, T.-T. Tseng, H.-R. Ke, and C.-T. Sun, "Social trend tracking by time series based social tagging clustering," *Expert Syst. Appl.*, vol. 38, pp. 12 807–12 817, 09 2011.

[27] K. Zhao, M. A. Wulder, T. Hu, R. Bright, Q. Wu, H. Qin, Y. Li, E. Toman, B. Mallick, X. Zhang, and M. Brown, "Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A bayesian ensemble algorithm," *Remote Sensing of Environment*, vol. 232, p. 111181, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425719301853

[28] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – challenges in topic discovery, data collection, and data preparation," *International Journal of Information Management*, vol. 39, pp. 156 – 168, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401217308526

[29] L. Aiello, G. Petkos, C. Martín Dancausa, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on Multimedia*, vol. 15, pp. 1–1, 10 2013.

[30] R. Diwate, "Data mining techniques in association rule : A review," 01 2014.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: Tf-idf approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, March 2016, pp. 61–66.

[36] V. Gurusamy and S. Kannan, "Preprocessing techniques for text mining," 10 2014.

[37] K. Chekima and R. Alfred, "An automatic construction of malay stop words based on aggregation method," in *Soft Computing in Data Science*, M. W. Berry, A. Hj. Mohamed, and B. W. Yap, Eds. Singapore: Springer Singapore, 2016, pp. 180–189.

[38] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, pp. 693–723, 2007, cited By 195. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-33947615175&partnerID=40&md5=6b15c4071322fe1495a23c8412a1d4fd

[39] M. Elias Polus and T. Abbas, "Development for performance of porter stemmer algorithm," *Eastern-European Journal of Enterprise Technologies*, vol. 1, no. 2 (109), p. 6–13, Feb. 2021. [Online]. Available: http://journals.uran.ua/eejet/article/view/225362

[40] T. Xia and Y. Chai, "An improvement to tf-idf: Term distribution based term weight algorithm," *JSW*, vol. 6, pp. 413–420, 03 2011.

[41] Y. Lei, "4 - clustering algorithm–based fault diagnosis," in *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery*, Y. Lei, Ed. Butterworth-Heinemann, 2017, pp. 175 – 229. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128115343000044

[42] A. Reynolds, G. Richards, B. Iglesia, and V. Rayward-Smith, "Clustering rules: A comparison of partitioning and hierarchical clustering algorithms," *J. Math. Model. Algorithms*, vol. 5, pp. 475–504, 12 2006.

[43] 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.

[44] 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.

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

[46] V. N. Gudivada, D. L. Rao, and A. R. Gudivada, "Chapter 11 - information retrieval: Concepts, models, and systems," in *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, ser. Handbook of Statistics, V. N. Gudivada and C. Rao, Eds. Elsevier, 2018, vol. 38, pp. 331 – 401. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169716118300245

[47] M. Melucci, *Vector-Space Model*. Boston, MA: Springer US, 2009, pp. 3259–3263. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_918

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.

Loh Boon Jing is a Computer Science graduate from Faculty of Computing and Informatics, Universiti Malaysia Sabah in 2017.

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.

Raymond Alfred is the Chief Executive Officer of Allyssa Research Group. He received his bachelor's degree in Surveying from University of East London, United Kingdom. He obtained his Ph.D. in Ecological Processes and Modelling from the Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia. His areas of expertise include Spatial Analysis, Geographic Information System and Solutions, Carbon and Biodiversity Mapping in Forest-Oil Palm Lanscape