INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Indonesian Online News Extraction and Clustering Using Evolving Clustering

Muhammad Alfian[a,*], Ali Ridho Barakbah[a], Idris Winarno[a]

[a] *Informatics and Computer Engineering Department, Politeknik Elektronika Negeri Surabaya, Indonesia*
Corresponding author: [*]ini.muhalfian@gmail.com

***Abstract*—43,000 online media outlets in Indonesia publish at least one to two stories every hour. The amount of information exceeds human processing capacity, resulting in several impacts for humans, such as confusion and psychological pressure. This study proposes the Evolving Clustering method that continually adapts existing model knowledge in the real, ever-evolving environment without re-clustering the data. This study also proposes feature extraction with vector space-based stemming features to improve Indonesian language stemming. The application of the system consists of seven stages, (1) Data Acquisition, (2) Data Pipeline, (3) Keyword Feature Extraction, (4) Data Aggregation, (5) Predefined Cluster using Automatic Clustering algorithm, (6) Evolving Clustering, and (7) News Clustering Result. The experimental results show that Automatic Clustering generated 388 clusters as predefined clusters from 3.000 news. One of them is the unknown cluster. Evolving clustering runs for two days to cluster the news by streaming, resulting in a total of 611 clusters. Evolving clustering goes well, both updating models and adding models. The performance of the Evolving Clustering algorithm is quite good, as evidenced by the cluster accuracy value of 88%. However, some clusters are not right. It should be re-evaluated in the keyword feature extraction process to extract the appropriate features for grouping. In the future, this method can be developed further by adding other functions, updating and adding to the model, and evaluating.**

***Keywords*—Evolving clustering; incremental clustering; news extraction; stemming.**

Manuscript received 12 Mar. 2021; revised 7 Jul. 2021; accepted 5 Aug. 2021. Date of publication 30 Sep. 2021.

I. INTRODUCTION

Online electronic media is the third generation of mass media after print and electronic media. The first mass media generation is print media, such as newspapers, magazines, and books. Meanwhile, the second mass media generation is electronic media, including radio, television, and film/video [1]. The mass media grows and develops overtimes. However, the change in the mass media generation has not eliminated the previous generation because the mass media are complementary in presenting the information. In the end, online media came to the third generation of mass media in various forms. Online media provides digital news that can be accessed anywhere and anytime from various gadgets such as computers or cell phones. Online media presents a new way of processing, producing, and disseminating news creates a new field for the media industry. Online media technology creates the most capable network compared to other mass media. Especially in terms of basic equipment, composing components, architecture, and various other capabilities [2]. Online media is growing increasingly massively in this digital era. The ease of setting up online media and low maintenance costs make online media grow rapidly. Based on data released by the Minister of Communication and Information Technology [3]. Online media in Indonesia reached 43,000 in 2018. Meanwhile, according to the Indonesian Journalists Association (PWI), online media circulating in Indonesia reached 47,000 media in 2019 [4]. This number increased 4,000 over one year. This number will continue to grow along with the development of the digital industry: the more online media, the more news is circulating in the community. Based on a survey conducted by Puspitasari *et al.* [3], every online media publishes at least one to two stories every hour. The amount of news will increase as online media grows. It exceeds the information processing capacity of the human brain. This has an impact on human mental health, resulting in confusion and psychological distress [5]. Therefore, humans cannot access the actual news effectively. Humans cannot read all the news on online media one by one because it will take much time. Online media produce news with the same information from one media to another. Every online media publishes news with almost similar headlines and has the same name entity [6]. Identifying news with similar

information looks easy for intelligent humans. However, it is not easy for the computer to accurately identify the same related discussions because news is classified as unstructured data. News must be changed from unstructured data to structured data.

One of the stages of the data transformation process is the text preprocessing stage. This stage plays an important role in information retrieval to extract relevant information [7]. Text preprocessing has several stages; one of them is stemming. Stemming is the process of converting affixed words into their original form (root word). The stemming process requires special attention because it is specific to the language in the text in its application. However, the stemming process in Indonesia still leaves some unresolved problems [8]. In general, the problem with the stemming process is over stemming or under stemming. For this reason, a dictionary-based approach can be an alternative in improving the Indonesian language stemming algorithm.

After the news has turned into structured data, the next process is to classify it according to the news discussion topic. News discussion topic is always changing according to the issues in society. For this reason, this study focuses on a clustering approach. The clustering approach can recognize the data pattern and group news without being tied to predefined labels. This makes the grouping approach model more dynamic and adaptable to developments in news topics. However, there are unresolved problems in the clustering algorithm. The most common problem is the need for traditional clustering to know how many clusters are required before clustering [9]. News discussion topics are always growing, and the number of them cannot be ascertained. In addition, traditional clustering algorithms can work well on limited datasets. However, this is not compatible with real-time requirements [10], especially in the news. The algorithm needs to reprocess all the data to get a new cluster. This is a time-consuming and computationally heavy process as the news data grows. For this reason, in this study, we propose the extraction of news features using the Vector Space approach in the stemming process. Then we cluster the news according to the information contained automatically and incrementally. Each resulting cluster will take a representative news item, making it easier for readers to understand the information in one cluster to make news reading activities more effective. There are several studies related to the process of stemming and news clustering.

A. Stemming process based on Confix Stripping

Research on stemming-lemmatization based on confix-stripping was initiated by Mirna Adriani and Bobby Nadzief [11]. This approach analyzes words from a morphological point of view. The research claims that the Nadzief Adriani algorithm is very accurate in converting affixed words into root words. However, there are still a few over stemming problems remaining. Over stemming is an incident where the algorithm trimmed too many words, thus losing meaning. Research conducted by Linggar [12] tried to prove the accuracy of the Nadzief Adriani algorithm. The study compared the Nadzief Adriani algorithm and Paice Husk. The accuracy obtained by the Nazief Adriani, and Paice Husk Algorithms is 91.87% and 64.43%. Meanwhile, the average processing time results for 200 sentence data for Nazief Adriani Algorithm 7.3516 / second and Paice Husk Algorithm 7,2510 / second. Thus, the algorithm with a better accuracy level is Nazief Adriani, but Paice Husk has a better processing time. Meanwhile, Rizki [13] summarized the research and then compared it among several stemming approaches. There are indications that the most accurate stemmer algorithm is not the only way to achieve the best performance in information retrieval (IR). In this study, seven Indonesian stemmer algorithms and English stemmer algorithms are compared. The approaches being compared include Nazief, Arifin, Fadillah, Asia, Enhanched Confix Stripping (ECS), Arifiyanti and Porter. The results show that Arifiyanti's development algorithm is the best algorithm for Indonesian text processing purposes with a score of 0.648.

Meanwhile, in another study, Simarangkir [14] tried to compare several stemming algorithms at once. This study uses two dictionary-based stemming algorithms and two stemming algorithms using affix rules. The results of the tests conducted show that the fastest stemming process is in the Vega algorithm. Meanwhile, the highest accuracy is in the Nazief and Adriani algorithms.

Several researchers tried to improve the Nadzief Adriani algorithm, including Asian [8], which tried to add a disallowed prefix rule. The results show an increase in accuracy from 93% to 95%. Meanwhile, Prihatin [15] also made improvements to the Nadzief Adriani algorithm by adding a dictionary and adding rules. This study claims that the accuracy of these improvements is increased to 94%. Until now, there is still no Indonesian stemming algorithm that has 100% accuracy.

B. Incremental News Clustering

The news clustering collects news from various sources. Then it groups news based on the proximity of its features. Research conducted by Puspitasari *et al.* [3] uses the Automatic Clustering algorithm to classify news automatically. The research has not shown optimal results because the time span for news collection is limited, so the news that can be grouped is also limited. In addition, the news data used is still static with a certain time span.

Another study that uses clustering is a study conducted by Sigita *et al.* [16], using the Online Clustering approach in processing news that grows over time. This Online Clustering Algorithm uses Vector Quantization (VQ) principles to create dynamic clusters. The results of this study obtained a precision of 70.9%. The remaining problems are that some news only have one keyword, thus affecting the accuracy of the clustering process. The Incremental Clustering method was also used by Azzopardi *et al.* [17] to group news into event-centric clusters. The clustering method adopts the Bisecting K-Means algorithm, which can run instantly (single pass) without cluster reorganization. This algorithm runs very fast, with an estimated processing time of approximately 1 second for one story. However, this algorithm works well only for specific news (discussing certain events), while it does not work well when general clustering news.

Meanwhile, the research conducted by Bakr [18] used an Incremental Density-based algorithm. It is introduced to build and update cluster datasets gradually. The experimental results show that the proposed algorithm has a significant increase in program execution time at runtime. This affects

the performance of the clustering algorithm. Apart from the Incremental Density-based algorithm, there are several algorithms related to news grouping. Research conducted by Laban [19] used a Community Detection algorithm approach. This algorithm is based on heuristics with modularity optimization. Meanwhile, another approach is carried out by Florence [20] using COH-K-means (Constrained Hierarchical K-Means).

II. Material and Method

This research proposes applying the Evolving Clustering method to group news incrementally on the Big Data platform. In previous studies, the clustering process did not run with incremental data, but static data were made incremental. We use the Big Data platform to ensure continuous data availability and run the clustering process with low latency. This allows the news to be grouped as soon as the news is crawled. Processing in big data environments is also possible with parallel computing, maximizing the computation process and making processing times faster. The Big Data platform used includes Apache Kafka as pipeline data and Apache Spark as data processing.

The source of the news is the RSS (Really Simple Syndication) of several online media. News from each RSS is passed into the database and then passed into the Keyword Feature Extraction stage. One of the processes in Keyword Feature Extraction is stemming. This study uses a new approach to the stemming process. We propose vector space-based stemming in this study. This is different from the previous approach, which used word morphology; vector space uses the dictionary model as a reference. The vector space model is the result of the Indonesian dictionary mapping process. This Vector Space Model is a reference in finding root words. Then the news goes into the Data Processing stage. At the Data Processing stage, the Evolving Clustering algorithm is applied. This algorithm classifies news according to its topic.

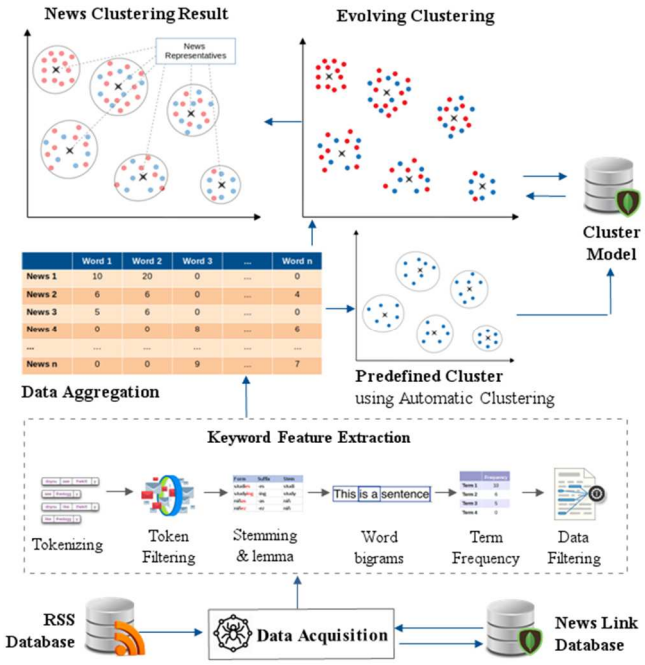

Fig. 1 System Design

From each news cluster, one news item is taken as representative news from each cluster. The representative news is displayed on the main page of the application. Thus, it reduces the number of news displayed. This research consists of six stages: Data Acquisition, Keyword Feature Extraction, Data Aggregation, Predefined Cluster, Evolving Clustering, and News Clustering Result. The overall system design is shown in Fig 1.

A. Data Acquisition

The first stage of this research starts with data acquisition. At this stage, the news data is obtained from news RSS. This study took news data from 62 RSS links from 17 online media that had been conducted by previous studies [21]. In addition, we added 18 additional RSS links from two online media to increase the diversity of the news. Table I. shows examples of RSS links in use.

TABLE I
RSS Link Example

Num	Name	Link
1.	Viva News	http://rss.viva.co.id/get/all
2.	Okezone	http://sindikasi.okezone.com/index.php/rss/0/RSS2.0

RSS contains the latest news from each online media. Therefore, the crawler program is designed to run every five minutes to get the latest news from each online media. The program will check the database, so there is no duplication of data. The crawler program will scrape the link to get the news attribute if the news data has never been retrieved before, as shown in Table II. The attribute data obtained include a link, source, description, publish date, image, title, and text. After the news attribute data is obtained, the crawler program sends the data to the database.

TABLE II
News Attribute Example

Attribute	Value
Link	*https://bola.okezone.com/read/2020/06/17/45/2231678/man-city-vs-arsenal-guardiola-lakukan-rotasi-pemain-di-sisa-liga-inggris-2019-2020*
Source	*http://sindikasi.okezone.com/index.php/rss/0/RSS2.0*
Description	*Guardiola akan melakukan rotasi pemain mengingat padatnya jadwal sisa Liga Inggris 2019-2020.*
Image	*https://img.okezone.com/dynamic/content/2020/06/17/45/2231678/man-city-vs-arsenal-guardiola-lakukan-rotasi-pemain-di-sisa-liga-inggris-2019-2020-QeinYlNFfF.jpg?w=300*
Publish date	*17/06/2020 16:05:05*
Title	*Man City vs Arsenal, Guardiola Lakukan Rotasi Pemain di Sisa Liga Inggris 2019-2020*
Text	*MANCHESTER – Pelatih Manchester City, Josep Guardiola, mengaku akan melakukan rotasi pemain di sisa Liga Inggris 2019-2020. Rotasi pemain dilakukan Guardiola agar kondisi fisik Sergio Aguero dan kawan-kawan dalam kondisi terbaik ...*

B. Keyword Feature Extraction

Keyword feature extraction consists of two main processes, *i.e.,* data cleaning and keyword selection. The two processes are broken down into the following six steps:

1) Tokenizing: Clear characters other than the alphabet (A-Z) in the text, then split into a series of words.

2) Token Filtering: Remove stop-words from the series of words.

3) Stemming & Lemma: Transform word into the root word by removing affixes and refinement into a basic form.

4) Word Bigram: Automatically create word phrases with two-word composition without neglecting the preceding series of words

5) Term Frequency: Count the number of words or phrases (w_i) in the document (d) then transformed into word frequency matrix (F_d).

$$F_d = [\, f_{w_i d} \quad \cdots \quad f_{w_n d} \,],\ w \in d \tag{1}$$

6) Data Filtering: Select words as keywords by filtering words by their frequency. We set the filtering threshold (th_d) at half the maximum word frequency as shown in equation (2). Then, as shown in equation (3), we apply a filtering threshold (th_d) to the term frequency (F_d) to reduce the number of terms, and then it referred to as keywords.

$$th_d = \frac{1}{2}\, max\{F_d\} \tag{2}$$

$$F'_d = \{\, f_{w_i d} \mid f_{w_i d} \geq \mathrm{th_d} \,\},\ w \in d \tag{3}$$

C. Stemming using Vector Space

Stemming and Lemmatization is the process of removing affixes and converting them into root words. An example is the word "*perkataan*". The word "*perkataan*" is removed from the affix "*per-*" and "*-an*" so that it turns into "*kata*" as the root word. Another example is the word "*meninggal*" is removed from the affix "*me-*"so that it turns into "*ninggal*". Then the word "*ninggal*" is converted into the root word, *i.e.,* "*tinggal*". Our proposed stemming method consists of two main processes, *i.e.,* building Vector Space Model and Stemming using Vector Space as shown in Fig 2.

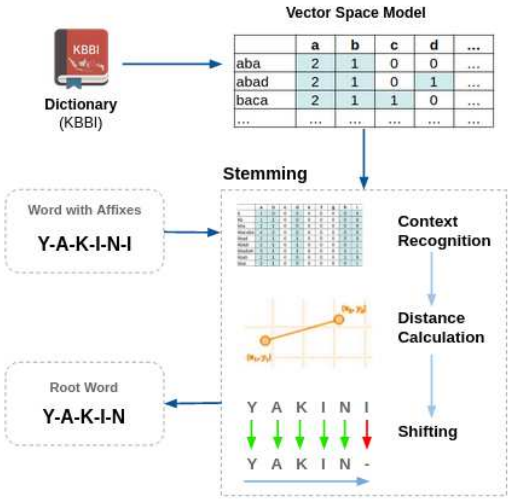

Fig. 2 Stemming using Vector Space Illustration

The first process is building Vector Space Model. Initially, we collect the root word (W) taken from the *Kamus Besar Bahasa Indonesia* (KBBI) as shown in equation (5). In addition, we also list alphabet (A) from letter A to Z as shown in equation (4). Then, we break down the words into the alphabet that makes them up and count their frequency ($f_{a_j w_i}$). If the word does not consist of a specific alphabet, it returns zero (0). This function is shown in equation (6). Then the results of equation (4) are stacked into alphabet matrix form word (V_{w_i}) and stack them into matrices (V) as shown in equations (7) and equation (8). This matrix is called the Vector Space Model (VSM).

$$A = \{\, a \mid a \in alphabet \,\} \tag{4}$$

$$W = \{\, w_i \mid 1 \leq i \leq n \,\},\ n > 0 \tag{5}$$

$$v_{w_i a_j} = \begin{cases} f_{a_j w_i}, & a \in w_i \\ 0, & a \notin w_i \end{cases} \tag{6}$$

$$V_{w_i} = [\, v_{w_i a_j} \quad \cdots \quad v_{w_i a_m} \,],\ m = \dim(A) \tag{7}$$

$$V = \begin{bmatrix} V_{w_i} \\ \vdots \\ V_{w_n} \end{bmatrix} \tag{8}$$

After the model is created, the next step is stemming. The stemming process consists of three main steps, *i.e.,* context recognition, distance calculation, and shifting. First, we must have a word with affixes (w_f) as the word to convert. Then, we break the word with affixes into the alphabet and convert them into a vector as in the previous step in equation (7). Then we calculate the distance (d) between the word with affixes (V_{w_f}) to the words in Vector Space Model (V_{w_i}) using Euclidean distance shown in equation (11). However, we did not use all the features to compare. We filter specific features using context recognition. This method reduces the number of features by comparing alphabet (A) that only exist in words with affixes (w_f) as shown in equation (9).

$$A' = \{\, a' \mid a' \in w_f \,\}, m = \dim(A') \tag{9}$$

$$d\left(X_{w_f}, V_{w_i}\right) = \sqrt{\sum_{j=0}^{m} (V_{w_i a'_j} - V_{w_f a'_j})^2} \tag{11}$$

Afterward, we get the ten closest words distance from the model (V) as the candidates (w_i) from the root word. Each candidate is converted into a vector of letter (Y). Moreover, the word with affixes is also converted into a letter (X) vector, as shown in equation (12). We checked the similarities between two words using shifting. This method is a simple check by counting the letter differences between two words. If there are equal letter, it returns 0. Otherwise, it returns 1 as shown in equation (13). This process is repeated for every letter in vector (m) and as much as the number of letter differences between the two vectors (n) as shown in equation (12). After that, we get minimum shifting value (s_{min}). The word with the minimum shifting value is the root word, as shown in equation (15).

$$X = \{\, x \mid x \in w_f \,\},\quad Y = \{\, y \mid y \in w_i \,\} \tag{12}$$

$$f(x_i, y_i) = \begin{cases} 0, & if\ x_i = y_i \\ 1, & if\ x_i \neq y_i \end{cases} \quad (13)$$

$$m = \dim(X),\ n = m - \dim(Y) \quad (14)$$

$$s_{min} = \min_{j \in [0,n)} \{ \Sigma_{i=0}^{m} f(x_{i+j}, y_i) \} \quad (15)$$

D. Data Aggregation

This process aggregates the term frequency of each document (F'_{d_j}) to a large matrix (M) vertically, as shown in equation (17). However, before we aggregate all term frequencies of each document (F'_{d_j}), we aggregate words of all documents as shown in equation (16). If the word is not in a document, it is left zero. In short, we aggregate the data horizontally based on words and vertically based on documents. Table III shows an example of data aggregation.

$$F'_{d_j} = [\, f_{w_i d_j} \quad \cdots \quad f_{w_n d_j} \,],\ w \in \cup_{j=0}^{m} d_j \quad (16)$$

$$M = \begin{bmatrix} F'_{d_j} \\ \vdots \\ F'_{d_m} \end{bmatrix}, m > 0 \quad (17)$$

TABLE III
DATA AGGREGATION EXAMPLE

	Word-1	Word-2	Word-3	Word-n
News-1	5	15	0	…
News-2	0	0	23	…
News-3	0	0	0	…
News-n	…	…	…	…

E. Predefined Cluster

A predefined cluster is a cluster generated from initial documents that have been predefined. This stage is only run once as the initial model. We use the Automatic Clustering algorithm to create the cluster model automatically, without specifying the number of clusters. The Automatic Clustering developed by Barakbah [21] runs using the Single Linkage Hierarchical Algorithm with the agglomerative method. It automatically determines the number of clusters by finding the global optimum from the given data using the Valley Tracing method. An illustration of the Automatic Grouping algorithm is shown in Figure 3.

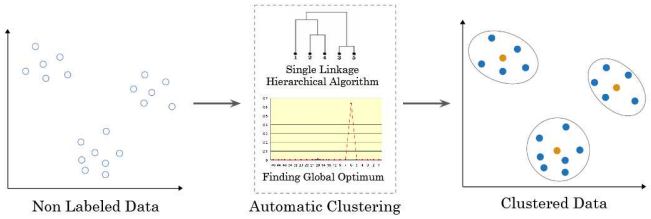

Fig. 3 Automatic Clustering Illustration

This Predefined Cluster stage produces a cluster model, as shown in Table IV. The attributes include cluster, radius, amount of data, and centroid. Radius is the distance from the centroid to the outer point of the cluster obtained using Euclidean distance. Member is the number of news registered in a cluster. Meanwhile, the centroid is the center point of the cluster.

TABLE IV
CLUSTER MODEL ATTRIBUTE EXAMPLE

Cluster	Radius	Member	Centroid
1	23.430	6	(9, 1.093), (24, 3.936), (25, 7.217), (30, 1.093),

The centroid (C_k) is formed from the average term frequency of cluster members as shown in equation (17).

$$C_k = [\ c_{kw_i} \quad \cdots \quad c_{kw_n}\ ],\ c_{kw_i} = \frac{\left(\Sigma_{j=0}^{m} f_{w_i d_j}\right)}{m} \quad (17)$$

F. Incremental Clustering

Incremental clustering is a clustering method that can continually adapt existing model knowledge when new data points are added. In this research, Automatic Clustering is used to produce the initial model because this method does not need to determine the number of clusters (determined automatically from the data pattern). The incremental clustering technically works when new data points emerge. It will determine which cluster is best suited for the new data points. The process of determining the cluster uses a competitive learning approach, *i.e.,* Learning Vector Quantization (LVQ) algorithm [22].

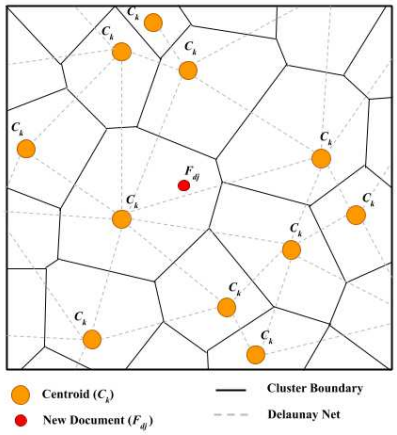

Fig. 4 Illustration of Cluster Model using Voronoi Diagram

LVQ is a type of Artificial Neural Network based on a prototype supervised learning classification algorithm and trained its network through a competitive learning algorithm similar to Self-Organizing Map. LVQ helps to find data structures by building class boundaries based on prototypes, a nearest-neighbor rule, and a winner-takes-it-all paradigm, as shown in Fig 4. This visualization uses Voronoi Diagram to make it easier to understand. It shows the arrangement of prototype points in a 2-dimensional plane. Each prototype point has its region called a cell/cluster. Each cluster has boundaries resulting from Delaunay Triangulation. We use the centroids as a prototype from the model because centroid is sufficient as a representation of the cluster

Figure 5 shows the Incremental Clustering Architecture. F_{d_j} is the vector conversion of the new data point. Each a new data point appears, we calculate the distance (d) from the new point to all prototype clusters/centroids (C_k) as shown in equation (18). This distance (d) indicates the degree of similarity between the new data point (F_{d_j}) and the prototype / centroid (C_k). The small distance means that the two points

have a high degree of similarity. It also means that the new data points (F_{d_j}) have a higher probability of being members of the prototype cluster (centroid). We select the most similar cluster (winning node) by obtaining the shortest distance as shown in equation (19). Then, the system will return the cluster number (k). After the system gets the right cluster (winning node), the system will learn. Learning is updating the weight of the centroid (winning node) by changing the position of the centroid (C'_k). We change the position of the centroid point using the standard competitive learning rule as shown in equation (20)

$$d\left(C_k, F_{d_j}\right) = \sqrt{\sum_{i=0}^{n}(c_{kw_i} - f_{w_i d_j})^2} \quad (18)$$

$$f\left(C_k, F_{d_j}\right) = \min_{j\in[0,m)}\{d\left(C_k, F_{d_j}\right)\} \quad (19)$$

$$c'_{kw_i} = c_{kw_i} + \alpha\left(f_{w_i d_j} - c_{kw_i}\right),\ \alpha \neq 0 \quad (20)$$

$$C'_k = [\ c'_{kw_i} \quad \dots \quad c'_{kw_n}\ ] \quad (21)$$

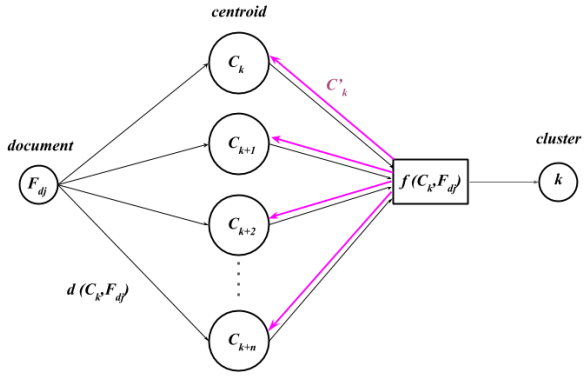

Fig. 5 Incremental Clustering Architecture

The incremental Clustering method does not need to regroup all data to a new model for each data growth. However, this method updates the model itself using LVQ system. So, this method really saves processing time. This method is effective in handling incremental data, but it is still not effective handling data in the real, ever-evolving environment. The effectiveness of the method depends on the initial model generated by Automatic Clustering. So, we also propose a function for incremental clustering to maintenance the model.

G. Evolving Clustering

Evolving clustering is an incremental clustering method with the ability to add models as the environment evolves. This method makes it possible to add a model by creating a new cluster if there is no proper cluster for the new data points. Moreover, we also propose the cluster boundary using a circle shape. We assume that the cluster has a circular shape regardless of its dimensions. Fig 6 shows the illustration of the Cluster Model using Evolving Clustering in a 2-dimensional plane.

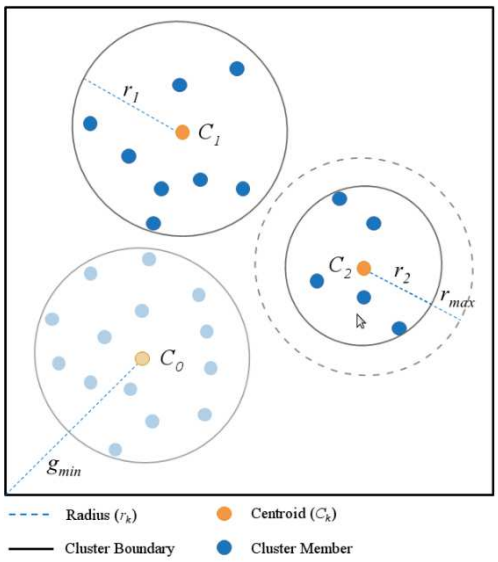

Fig. 6 Illustration of Cluster Model in Evolving Clustering

First, we define the unknown cluster. The unknown cluster is a cluster that has no meaning, is indicated by having the closest distance (g_{min}) from the centroid (C_k) to the zero point (O) as shown in equation (23). The unknown cluster is usually having the most number of members, because it is like a dump area. In further step, this cluster is ignored because it is not significant and can affect the accuracy of model.

$$d(C_k, O) = \sqrt{\sum_{i=0}^{n}(c_{kw_i} - o_{w_i})^2} \quad (22)$$

$$g_{min} = \min_{k\in[0,m)}\{d(C_k, O)\} \quad (23)$$

After that, we get the radius of cluster (r_k) by calculating the distance between the centroid and the outer data point in the cluster. Technically, we get the maximum distance from centroid (C_k) to all members (F_{d_j}) in the cluster as shown in equation (25). The cluster radius will be the identity of each cluster. However, it becomes a problem if a cluster has a small radius. Clusters with a small radius will not expand because they have limited space inside the radius. Therefore, we prefer to automatically make the general radius for all clusters by getting the maximum radius (r_{max}) from all cluster radius (r_k), except the unknown cluster as shown in equation (26). In the next step, we define the threshold (r_t) obtained from the maximum radius multiplied by a beta coefficient (β) as the general radius for all clusters as shown in equation (27).

$$d\left(C_k, F_{d_j}\right) = \sqrt{\sum_{i=0}^{n}(c_{kw_i} - f_{w_i d_j})^2} \quad (24)$$

$$r_k = \max_{k\in[0,m)}\{d\left(C_k, F_{d_j}\right)\} \quad (25)$$

$$r_{max} = \max_{k\in[0,n)}\{r_k\} \quad (26)$$

$$r_t = \beta\ .\ r_{max}, \beta > 0 \quad (27)$$

Evolving clustering has two conditions for dealing with new data points. First, if a new data point appears in a cluster, the new data point becomes a cluster member and the centroid position is updated. Second, if new data points appear outside of the entire cluster, then the new data points create a new

cluster. An illustration of the Incremental Clustering algorithm is shown in Figure 7.

The first condition is if new data points appear in a cluster. When new data points appear, we calculate the distance (d) between the new data points (F_{d_j}) and all cluster centroid (C_k). This is almost similar to Incremental Clustering. However, the Evolving Clustering determines the threshold (r_t) If the distance (d) is lower or equal to the threshold (r_t), the new data point becomes a cluster member. If no cluster meets the requirements, then the cluster enters the second condition. The second condition is that if new data points appear outside the entire cluster, then the new data points create a new cluster. The new cluster will have a radius equal to the value of the threshold (r_t) and it has only one member and becomes a centroid.

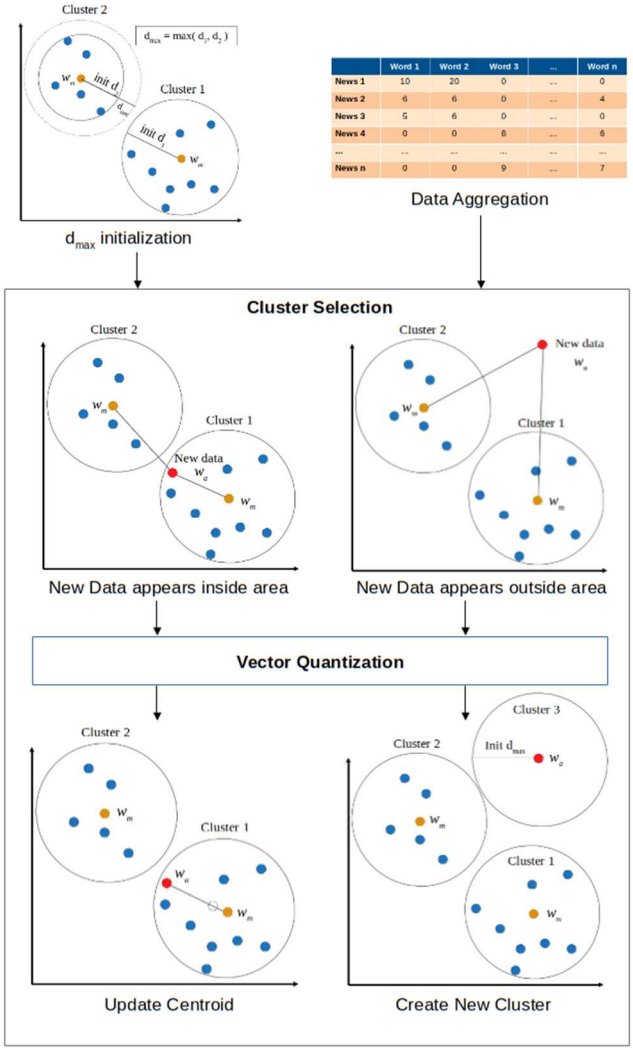

Fig. 7 Evolving Clustering Illustration

H. News Clustering Result

In this study, we propose a method to present item representative from a cluster. This method can reduce the number of items displayed and help to get significant items in a cluster. Technically, we got a representative item from an item that has the closest distance to the centroid. This item represents all members of the cluster because it has features that are almost similar to centroids. Meanwhile, the centroid is the center point of the cluster which displays a summary of all members. From the representative item, we calculate the distance to another item. The item that has the closest distance to the representative item has a higher probability of being similar to the representative item. The item that has the closest distance to the representative item will be called a related item. This item helps to describe the information displayed by the representative item. This item displays beside the representative item in a cluster. Meanwhile, another item in the cluster will be hidden. We use Euclidean Distance as the distance calculation. In this study, we used the news as items in the cluster.

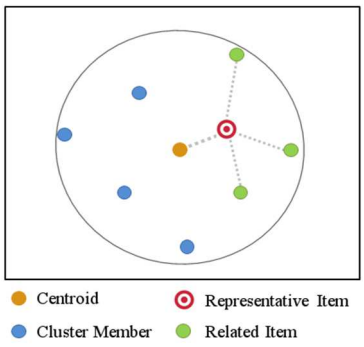

Fig. 8 Cluster Anatomy

III. Results and Discussion

The experiment consists of five parts, *i.e.,* Stemming using Vector Space, Predefined Cluster, Evolving Clustering, News Clustering Result, and Evaluation.

A. Stemming using Vector Space

Before carrying out the test, the first step is to collect Indonesian root word data as a model for building Vector Space-based stemming. The Indonesian root words are taken from KBBI. There are 28,526 root words obtained from KBBI. Then we map out the root words based on the alphabet that makes them up. For example, the word "*aba-aba*" consists of the letters A and B. There are four letters A and two letters B in composing the word "*aba-aba*". Meanwhile, the word "*abad*" consists of letters A, B, and D. Letter A consists of two letters and letters B and D have one letter. Figure 9 shows the results of a letter mapping called the Vector Space Model. This model will be used as a reference in the search for root words. Therefore this approach is called the Vector Space Model approach, because the model of this approach uses a dictionary that has been mapped based on letters. The dictionary mapping is based on letters in the form of Vector Space because one word can have 26 features according to its constituent letters.

	a	b	c	d	e	f	g	h	I
A	1	0	0	0	0	0	0	0	0
Ab	1	1	0	0	0	0	0	0	0
Aba	2	1	0	0	0	0	0	0	0
Aba-aba	4	2	0	0	0	0	0	0	0
Abad	2	1	0	1	0	0	0	0	0
Abadi	2	1	0	1	0	0	0	0	1
Abadiah	3	1	0	1	0	0	0	1	1
Abah	2	1	0	0	0	0	0	1	0
Abai	2	1	0	0	0	0	0	0	1

Fig. 9 Vector Space Model Result

Then the next step is to prepare the word with the affix as the target data. We process the target data one by one. Each target data was mapped into a vector space with 26 features (according to alphabet). Then we calculated the distance from the target data to all the root words in the Vector Space Model. We calculated distance with context recognition. Context recognition compares the target data to a vector space model in a particular feature. For example, suppose we have "*yakini*" target data. The word "*yakini*" consists of the letters A, I, K, N, and Y. So, we only counted the distances of the 5 letters, while the others were ignored. This is called context recognition. Table V shows the results of calculating distances using context recognition. In this step, we got the top-10 words closest to the target word as candidates of the root words.

TABLE V
DISTANCE CALCULATION USING CONTEXT RECOGNITION

Target word	Closest Word	Distance
	Linyak	0.0513
	Nyarik	0.0513
	Nyamik	0.0513
	Minyak	0.0513
	Yakni	0.0513
	Takyin	0.0513
Yakini	***Yakin***	**0.0513**
	Paniki	0.0645
	Rinyai	0.0645
	Kiani	0.0645

Then we compared the target data to the candidates of root word using shifting; we got the letter difference among them. The least difference value is the correct root word. Table VI shows the shifting value of word "*yakini*". From all candidates, the word "*yakin*" got a shifting value of 0. So, the root word for "*yakini*" is "*yakin*".

TABLE VI
SHIFTING VALUE CALCULATION

Target word	Closest Word	Distance
	Yakin	**0**
	Yakni	2
	Paniki	3
	Kiani	3
	Takyin	4
	Rinai	5
Yakini	*Linyak*	6
	Nyarik	6
	Nyamik	6
	Minyak	6

Then, we tested 498 words obtained from several articles. We measured the accuracy of the method and execution time of the program. The result is that our proposed method got an accuracy value of 65.66%. Meanwhile, the execution time is 0.363 seconds for each word. This is still quite slow, considering that one document consists of many words. The approach with Vector Space needs to be re-evaluated in order to get better results.

B. Predefined Cluster

In this study, we crawled news data from 78 RSS links between 21-22 April 2019. We managed to get 3,000 news data and used this data as a predefined cluster. We processed the data using the Automatic Clustering algorithm and obtained 388 clusters, including the unknown cluster. The unknown cluster consists of 2155 news. This means that the rest of the news is a member of 387 other clusters. Most clusters consist of 1 or 2 news. However, there is one cluster that contains the most news. It is cluster number 41, which discusses "the bombing in Sri Lanka". The topic of discussion regarding "the bombing in Sri Lanka" was indeed being discussed in the span of 2 days, so that it could form a cluster properly. Table VII shows the sample of the cluster formed in the Predefined Cluster.

TABLE VII
CLUSTER FORMED IN PREDEFINED CLUSTER

Cluster	News Headlines	Keyword
5	*Ketua DPRD Kulon Progo ajak perempuan berjuan di jalur politik* *Pejabat Perempuan di Pemkot Bandung Masih Minim* *Perempuan Penjaga Alam*	*Perempuan, juang, bandung, alam*
41	*PM Pakistan Kutuk Serangan di Sri Lanka* *Warganya Jadi Korban, Trump hingga Erdogan Kecam Teror Bom Sri Lanka* *Turki Kecam Keras Serangan Bom di Sri Lanka*	*Sri, lanka, srilanka, serang*
55	*Saingi Amazon dan Apple, Facebook Kembangkan Voice Assistant* *Facebook Tak Mau Kalah Bikin Asisten Suara yang Lebih Canggih* *Facebook Ungkap Jutaan Password Pengguna Instagram ke Karyawan*	*Facebook, asisten, password, guna*

C. Evolving Clustering

The first step in Incremental Evolving Clustering is determining the value of r_{max}.We determine the r_{max} value according to the formula that has been described. We get the maximum value of all cluster's radius (r_k) except the unknown cluster. In the experiment we did, we got a r_{max} value of 27.924. The r_{max} value is used as a threshold in determining the condition of the new incoming data.

We experimented using news data incrementally on May 6 - 8, 2019. We obtained 12.164 news. The cluster formed are 223 new clusters and 122 updated clusters, so that the entire cluster formed was 611 clusters. Table VIII. shows a sample of the updated news clusters from the old cluster. The new data point (news) is successfully grouped into clusters according to the keywords they have. One of them was cluster 41; talks about "the Bomb in Sri Lanka" continued so that news related to "the Bomb in Sri Lanka" would be included in cluster 41.

However, some clusters are not right, such as in cluster 5. Cluster 5 discusses "women". However, when viewed from the topics discussed, the news grouping was deemed inappropriate. This happens because humans can think contextually, which is built on knowledge. Meanwhile, this study classifies information retrieval-based news which refers

to keyword information attached to the news. This makes the resulting news grouping seem odd according to humans. However, in some cases, news clustering based on information retrieval can cluster news appropriately, as long as the news has sufficient keyword information to distinguish one news item from another.

TABLE VIII
UPDATED NEWS CLUSTER

Cluster	Status	News Headlines	Keyword
5	-	*Ketua DPRD Kulon Progo ajak perempuan berjuan di jalur politik*	Perempuan, juang, bandung, alam
	-	*Pejabat Perempuan di Pemkot Bandung Masih Minim*	
	-	*Perempuan Penjaga Alam*	
	New	*Parpol Diminta Prioritaskan Kaderisasi Kaum Perempuan*	
	New	*Menolak Tunduk pada Tradisi, Ini 3 Cerita Perempuan Myanmar, Korea dan China*	
	New	*Partai Politik Diminta Prioritaskan Kaderisasi Kaum Perempuan*	
41	-	*PM Pakistan Kutuk Serangan di Sri Lanka*	*Sri, lanka, srilanka, serang*
	-	*Warganya Jadi Korban, Trump hingga Erdogan Kecam Teror Bom Sri Lanka*	
	-	*Turki Kecam Keras Serangan Bom di Sri Lanka*	
	New	*Otoritas Sri Lanka Sebut Aset Teroris telah Dibekukan*	
	New	*Semua Tersangka Pengeboman Sri Lanka Telah Tertangkap Atau Tewas*	
	New	*Jumlah Wisatawan Asing di Sri Lanka Turun Pascabom*	
54	-	*Saingi Amazon dan Apple, Facebook Kembangkan Voice Assistant*	*Facebook, asisten, password, guna*
	-	*Facebook Tak Mau Kalah Bikin Asisten Suara yang Lebih Canggih*	
	-	*Facebook Ungkap Jutaan Password Pengguna Instagram ke Karyawan*	
	New	*WhatsApp akan Semakin Canggih, bisa Chat Messenger dan Instagram*	
	New	*Pengguna WhatsApp Bakal Bisa Chat ke Messenger dan Instagram*	
	New	*Unggah Hoaks di Instagram Bakal Disaring Pakai Alat Facebook*	

Meanwhile, the Evolving Clustering also generates new clusters in every news update. In the cluster model, no cluster discusses"*mudik*" ("homecoming"). Thus, when the Evolving Clustering method was running, a new cluster was formed with number 572, which discussed" mudik" ("homecoming"). This means that the Evolving Clustering method success in adding models as the environment evolves. The new clusters are shown in Table IX.

TABLE IX
NEW NEWS CLUSTER

Cluster	Status	News Headlines	Keyword
572	New	*3,7 Juta Warga Jabodetabek Mudik ke Jabar*	*Mudik*
	New	*Sebanyak 3,7 Juta Warga Jabodetabek Akan Mudik ke Jabar*	
	New	*Ini Tips Ikut Mudik Lebaran Gratis Bareng Jasa Raharja*	
	New	*Puncak Arus Mudik Diperkirakan 31 Mei 2019*	
	New	*Cek Jalur Mudik, Polda Jabar: Kondisi Siap Hadapi Lebaran*	
	New	Dishub Kabupaten Bandung Fasilitasi Mudik Gratis	

D. News Clustering Result

The next step is searching for the representative news from each cluster. The representative news is the news that has the closest distance to the centroid. We calculated the distance using the Euclidean distance. As shown in Table 10, the representative news from cluster 41 is "*Otoritas Sri Lanka Sebut Aset Teroris telah Dibekukan*" (Sri Lankan Authority Call Terrorist Assets Freeze). Then we got five news related to representative news, as shown in the table. The related news is ordered based on the distance between the news to the representative news.

TABLE X
NEWS CLUSTERING RESULT

Cluster	Status	News Headlines
41	Representative News	*Otoritas Sri Lanka Sebut Aset Teroris telah Dibekukan*
	Related News	*Pasca Bom Sri Lanka, Perdagangan Bursa Kolombo Libur*
		Bom di Sri Lanka, Tujuh Orang Ditangkap
		Semua Tersangka Pengeboman Sri Lanka Telah Tertangkap Atau Tewas
		Teror Bom di Sri Lanka, Netizen Ramai-Ramai Kirimkan Doa
		Jumlah Wisatawan Asing di Sri Lanka Turun Pascabom

The news clustering result is displayed on a web-based application, as shown in Fig 10. One application of the clustering method is to generate headline news. We choose one of the hot topics about "*Covid*" as shown image below. There is a lot of news in the cluster about "*Covid*", but we just show a few of them using the news representative method. The representative news is the big one. It was deliberately made bigger than the others to grab user's attention. Meanwhile, there are several related news besides the representative news. The related news is displayed sequentially based on the distance to the representative news.

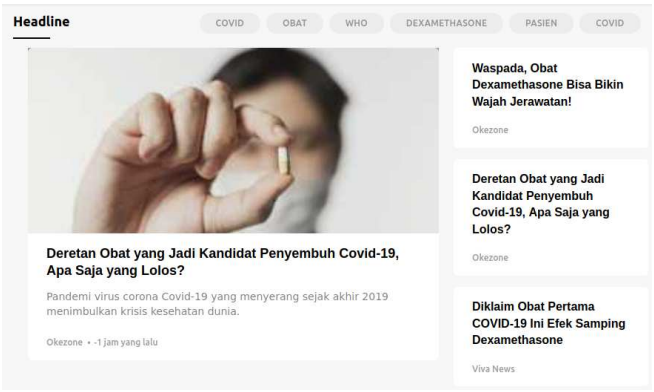

Fig. 10 News Representative from Cluster about Covid

E. Evaluation

Observations evaluated the Evolving Clustering methods. We made observations on each cluster produced on May 6 - 8, 2019. We got 100 cluster samples randomly from the whole cluster. Half of them is updated from the previous cluster, and half are new clusters that have never been formed before. We analyze news stories that form clusters that are similar or not to other news. If one piece of news does not fit into a cluster, then the cluster is considered wrong. Conversely, if the cluster has similar news, then the cluster is considered correct. Table XI shows the evaluation of the cluster.

TABLE XI
CLUSTER EVALUATION

Cluster	News Headlines	Keyword	Evaluation
5	*Ketua DPRD Kulon Progo ajak perempuan berjuan di jalur politik* *Pejabat Perempuan di Pemkot Bandung Masih Minim* *Perempuan Penjaga Alam* *Parpol Diminta Prioritaskan Kaderisasi Kaum Perempuan* *Menolak Tunduk pada Tradisi, Ini 3 Cerita Perempuan Myanmar, Korea dan China* *Partai Politik Diminta Prioritaskan Kaderisasi Kaum Perempuan*	*Perempuan, juang, bandung, alam*	*False*
41	*PM Pakistan Kutuk Serangan di Sri Lanka* *Warganya Jadi Korban, Trump hingga Erdogan Kecam Teror Bom Sri Lanka* *Turki Kecam Keras Serangan Bom di Sri Lanka* *Otoritas Sri Lanka Sebut Aset Teroris telah Dibekukan* *Semua Tersangka Pengeboman Sri Lanka Telah Tertangkap Atau Tewas* *Jumlah Wisatawan Asing di Sri Lanka Turun Pascabom*	*Sri, lanka, srilanka, serang*	*True*
54	*Saingi Amazon dan Apple, Facebook Kembangkan Voice Assistant* *Facebook Tak Mau Kalah Bikin Asisten Suara yang Lebih Canggih* *Facebook Ungkap Jutaan Password Pengguna Instagram ke Karyawan* *WhatsApp akan Semakin Canggih, bisa Chat Messenger dan Instagram* *Pengguna WhatsApp Bakal Bisa Chat ke Messenger dan Instagram* *Unggah Hoaks di Instagram Bakal Disaring Pakai Alat Facebook*	*Facebook, asisten, password, guna*	*True*

The results of the observations stated that 12 out of 100 clusters were wrong clusters. This proves that the Evolving Clustering algorithm has an accuracy of 88%. The evaluation also stated that most of the faulty clusters resulted from the cluster update process. The cluster update process still includes news not relevant to the previous news due to the lack of cluster selection in entering new data. Meanwhile, creating a new cluster has a few errors because this process creates small but many clusters so that the discussion of news is in one specific cluster.

IV. CONCLUSION

This study creates a representative news application using Evolving Clustering that runs in a Big Data environment. The application of the system consists of six stages, *i.e.* (1) Data Acquisition, (2) Keyword Feature Extraction, (3) Data Aggregation, (4) Predefined Cluster using Automatic Clustering algorithm, (5) Evolving Clustering, and (6) News Clustering Result. We get news data via 78 RSS links. The data acquisition process obtained 3,000 news items which were then processed with keyword feature extraction to produce keywords. Our approach using Vector Space in stemming stage has an accuracy value of 65.66%. The accuracy is still not optimal. However, this stemming method is a new approach, and it opens up opportunities for us to develop this method even better. The keywords are then aggregated into one large matrix and then processed using Automatic Clustering to produce 388 clusters as predefined clusters. Then the incremental data is processed using the Evolving Clustering algorithm. The Evolving Clustering Algorithm produces 223 new clusters and 122 updated clusters so that the total cluster formed is 611 clusters. Evolving clustering runs well, both updating model and adding the model. The performance of the Evolving Clustering algorithm is quite good, as evidenced by the cluster

accuracy value of 88%. However, several clusters are not quite right. It should be re-evaluated in the keyword feature extraction process to extract appropriate features for clustering. In the future, this method can be developed further by adding other functions, updates and add models, and several other functions to evaluate the model periodically.

NOMENCLATURE

w	notation for a word
d	notation for a document
$f_{w_i d}$	frequency word-i in a document
F_d	feature vector of document
th_d	threshold value in a document
a	notation for alphabet letter
$f_{a_j w_i}$	frequency of alphabet-j in the word-i
s_{min}	minimum shifting value
C_k	feature vector of centroid-k
O	vector of zero value
g_{min}	minimum distance value to ground (zero)
r_k	radius of cluster-k
α	coefficient for updating centroid
β	coefficient for determining threshold radius

REFERENCES

[1] A. S. M. Romli, *Jurnalistik Online: Panduan Mengelola Media Online*, I. Kurniawan, Bandung, Indonesia: Nuansa, 2012.

[2] S. S. Kurnia, *Jurnalisme Kontemporer*, Jakarta, Indonesia: Yayasan Pustaka Obor Indonesia, 2017.

[3] D. Z. E. Puspitasari, A. R. Barakbah, I. Winarno, "Automatic Representative News Generation using Automatic Clustering", in *Industrial Electronics Seminar (IES) 2012*, 2012.

[4] AMSI. (2019) Dari 47 Ribu, Baru 2.700 Media Online Terverifikasi Dewan Pers. [Online]. Available: https://www.amsi.or.id/dari-47-ribu-baru-2-700-media-online-terverifikasi-dewan-pers/

[5] J. B. Schmitt, C. A. Debbelt, F. M. Schneider, "Too much information? Predictors of information overload in the context of online news exposure", *Information Communication and Society*, vol. 21, no. 8, pp. 1151-1167, Apr. 2017.

[6] I. Subašić, B. Berendt, "Peddling or Creating? Investigating the Role of Twitter in News Reporting",in *Proc. ECIR 2011*, 2011, p 207-213, 2011.

[7] P. Virmani, S. Taneja, "A Text Preprocessing Approach for Efficacious Information Retrieval," *Smart Innovations in Communication and Computational Sciences*, pp.13-22, Jan. 2019.

[8] J. Asian, H. E. Wiliams dan S. M. M. Tahaghoghi, "Stemming Indonesian", in *Proceedings ACSC '05,* 2005, p. 307-314.

[9] Z. Aliniya, S. A. Mirroshandel, "A novel combinatorial merge-split approach for automatic clustering using imperialist competitive algorithm", *Expert System with Applications*, vol. 117, p. 243-266, Mar. 2019.

[10] M. K. Islam, M. M. Ahmed, K. Z. Zamli, "A buffer-based online clustering for evolving data stream", *Information Sciences*, vol. 489, p. 113-135, Jul. 2019.

[11] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi dan H. E. Wiliam, "Stemming Indonesian: A confix-Stripping Approach", *ACM Transactions on Asian Language Information Processing*, vol. 6, no. 4, Dec. 2007.

[12] L. D. Pratiwi, "Perbandingan Algoritma Nazief Adriani dan Paice Husk untuk Proses Stemming Teks Bahasa Indonesia," B.Sc thesis, UIN Sunan Gunung Djati, Bandung, Indonesia, Oct. 2019.

[13] A. S. Rizki, A. Tjahyanto dan R. Trialih, "Comparison of Stemming Algorithm on Indonesian Text Processing", *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 17, no. 1, pp. 95-102, Feb. 2019.

[14] M. S. H. Simarangkir, "Studi Perbandingan Algoritma-Algoritma Stemming untuk Dokumen Teks Bahasa Indonesia," *Jurnal INKOFAR,* vol. 1, no. 1, Jul. 2017.

[15] P. Prihatin, I. D. Putra, I. Giriantri dan M. Sudarma, "Stemming Algorithm for Indonesian Digital News Text Processing", *International Journal of Engineering and Emerging Technology*, vol. 2, no. 2, pp. 1-7, Mar. 2018.

[16] M. Sigita, A. R. Barakbah, E. M. Kusumaningtyas, I. W., "Automatic Representative News Generation using On-Line Clustering", *EMITTER International Journal of Engineering Technology*, vol. 1, no. 1, pp. 107-113, Dec. 2013.

[17] J. Azzopardi, C. Staff, "Incremental Clustering of News Reports", *Algorithms - Open Access Journal, v*ol. 5, no. 3, pp. 364 - 378, Dec. 2012.

[18] A. M. Bakr, N. M. Ghanem dan M. A. Ismail, "Efficient Incremental Density-based algorithm for clustering large datasets," *Alexandria Engineering Journal,* vol. 54, no. 4, pp. 1147-1154, Dec. 2015.

[19] P. Laban dan M. Hearst, "newsLens: building and visualizing long-ranging news stories," in *Proceedings of the Events and Stories in the News Workshop*, 2017, p. 1-9.

[20] R. Florence, B. Nogueira dan R. Marcacini, "Constrained Hierarchical Clusteriing of News Events," in *Proceedings of the 21st International Database Engineering & Applications Symposium (IDEAS) 2017*, 2017, p. 49-56.

[21] I. Shabirin, "Cluster Based News Representative Generation with Automatic Incremental Clustering", M.Eng. thesis, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia, Jul. 2017.

[22] A. R. Barakbah, K. Arai, "Reversed Pattern of Moving Variance for Accelerating Automatic Clustering", *EEPIS Journal*, vol. 2, no. 9, pp. 15-21, 2004.

[23] T. Kohonen, *Self-Organizing Maps: Learning Vector Quantization*, ser. Springer Series in Information Sciences, Springer, Berlin, Heidelberg : Springer, 1995, vol 30.