A Comparison of Clustering Algorithm Specifying in Topical Cluster Tweets of the Theme of Ambon Tourism

Paul Rio Pelupessy

Faculty of Information Technology, Master of Information Systems, Satya Wacana Christian University, Salatiga, Indonesia

riopaul545@gmail.com

ABSTRACT

Twitter as a Social Medium under The Indonesian Social Media is nowadays commonly used as a source of data for research. With a library like Twitter4J, it is possible to access a user's tweets according to the user's needs. The most widely used research using Twitter data is grouping the user tweets based on a certain topic using the clustering method. This study aims at comparing the clustering algorithm to classify a tweet with a theme of the tourism of Ambon. The comparable algorithms are Lingo, Suffix Tree Clustering (STC): K-Means and K-Medoids. To compare these algorithms this study developed an application to Javanese language. The comparable aspects are purity, accuracy, precision, recall, and processing speed. A test result using 1000 tweet showed that the STC is an algorithm that has the fastest average processing time, which is 1.6161 seconds. While on the purity aspect, Lingo and STC algorithms have the same purity level, which is 1 with a level of accuracy above 80%. The study concluded that the STC algorithm is the best algorithm in the clustering tweet with the theme of Ambon Tourism.

Keywords: *Strategic Planning, TOGAF ADM, Enterprise Architecture Score Card*.

1. INTRODUCTION

Twitter has recently been used not only as a social media for people to interact digitally, but it can be used as a source of carrying data from certain research. Twitter provides features of Application Programmable Interface API) which allow software developers to access twitter data. By utilizing Twitter API, for example Twitter 4J, the tweet-tweet activities can be accessed as needed. The process of making tweets that adjust to the needs is done through filtering the tweets that can be accessed. Tweets filtering process can be done based on several aspects like geolocation, certain key words, tweets from certain accounts and etc.

The access twitter data can be processed, so it can produce various information according to user requirements, for example, information about topics that are being discussed by Twitter users. The resulting topics can be focused on certain themes such as politics, social, tourism from certain region.

The method that can be used in generating topics from Twitter data is the clustering method. The clustering method is a method for grouping documents or data into certain groups based on the similarity of the contents of the documents, where each group is labeled. The clustering method is used when it is not known how the data or documents should be grouped. Unlike the classification method in which determining classification is done before grouping the documents or data using training data. Thus, the clustering method can be used to extract certain information from a set of tweets, for example the topic discussed in the set of tweets. Besides generating the topics discussed in each tweet, by applying clustering method, the weight or percentage of each topic from a set of tweets can be determined.

In general, clustering method can be classified into two groups, they are clustering method with phrase-based and term-based approach. For example, the phrase-based clustering algorithm is Lingo and Suffix Tree Clustering (STC), while term-based clustering belongs to the K-Means and K-Medoids algorithms. Each of these algorithms has advantages and disadvantages. In the literature review, various studies that use twitter data for various purposes, but carrying research about the comparison of the above algorithms has not been done. Therefore, this study focused on the comparison of these algorithms like processing time, purity, precision and recall, and accuracy. Twitter data that used as test data is focused on Twitter data themed Ambon tourism.

2. LITERARY STUDY

The following are a number of researches which have been executed to raise a topical theme twitter or tweet:
Andonovski & Chorbev (2014) in a study titled “Clustering of tweets from popular athletes”: implement an algorithm to define the most popular topics of famous athletes. The algorithm uses Lucene to determine the most often used words in data sets and posts which used those words [1]. Ashktorab et al (2014) in a study titled “Tweetdr: Mining Twitter to inform Disaster Response” developed a Twitter-Mining tool that can extract information related to the actions of the disaster relief workers during natural disasters. The Tweetdr flow consists of three main parts: Classification, Clustering and Extraction. At the classification stage, SLDA method, SVM and Logistic Regression are used to identify tweets that report the victim and damage. In the clustering phase, filters are used to combine similar tweets, and in the extraction phase, token and phrase extractions are reported, which report specific information about classes of infrastructure damage, type of damage and casualties. This tool has been validated using tweets collected from 12 different kriris in the USA since 2006 [2].

Still a touch with natural disasters, in the year 2015 Saravanou & Valkanas (2015) executed research titled “Twitter Floods When it Rains: A Case Study of the UK Floods in early 2014”, in the study Saravanou & Valkanas (2015) explored the use of twitter for the mechanisms used in disaster relief and public security.

Specifically, research with case studies on floods that occurred during January 2014 in the UK and how it is depicted on twitter, based on tweets submitted by users. Saravanou & Valkanas (2015) presents a systematic algorithmic analysis of tweets collected in accordance with the use case scenarios provided by visual analysis tools. The goal is to identify more meaningful and effective ways to benefit from the abundance of twitter data in crisis management [3].

Sechelea et al (2015) in a study titled “Twitter Data Clustering and Visualization” proposed a system for acquiring, analyzing and visualizing twitter data. Twitter messages are accessed and stored in a distributed cluster and data is processed using algorithms implemented in the Map Reduce framework. Sechelea et al (2015) proposed an algorithm that was able to identify the main topic contained in a set of tweet data. The study also designed a method of visualization that allows to follow the intensity of twitter activity based on geolocation [4].

Godfrey et al (2015) in a study titled “A Case Study in Text Mining: Interpreting Twitter Data from World Cup Tweets” performed an analysis of 30,000 tweets collected from twitter before the 2014 World Cup began. Of the many tweets, there are many tweets that are not relevant to the World Cup theme. To overcome this DBSCAN algorithm and consensus matrix are used. Then K-Means and Non-Negative Matrix Factorization (NMF) algorithms are used to get the topic from the tweetsand compare the results [5].

Friedemann (2015) in a study titled clustering a Customer Base using twitter Data” carried out a research to cluster customers from a company by using data from twitter. The application included customizing ad campaigns, localizing the untouched market segments, and projecting sale trends. An outline of the technique used is to look for slices among twitter data which can be accessed by the public with built-in features [6].

Unnisa et al (2017) in a study titled “Opinion Mining on Twitter data using Unsupervised Learning Technique” examines the grouping of tweets into positive and negative clusters. In prior research, supervised machine learning such as Naïve Bayes and Suppor Vector Machine are used, in this research unsupervised learning approach is used as spectral clustering. Clustering results are also visualized using graphs [7].

Alnajran et al (2017) in a study titled “Cluster Analysis of Twitter Data: A Review of Algorithms” carried out a comparison analysis of the unsupervised learning approach used to cluster the twitter data. This comparison includes clustering methods, algorithms, cluster number, data size, distance size, clustering features, evaluation methods, and results obtained. The result of the study concluded that the unsupervised learning approach in digging data from twitter has several weaknesses [8].

3. RESEARCH METHOD

The developed model used at tweet clustering process about Ambon theme in this research can be seen in picture 1. There were four steps in this model, they are tweet access, preprocessing, clustering, and evaluation. The aim of doing tweet access is to tweet all themes of tourism in Ambon, while preprocessing tweet has aim to prepare the process of the tweet. There were two different block approaches in clustering step, they are phrase-based approach with Lingo and STC algorithm. The framework Carrot2 was used in this phrase approach. In term-based approach, the framework Java Machine Learning Java ML) was used with K-Means and K-Medois algorithms. The result of clustering from all algorithms was evaluated.

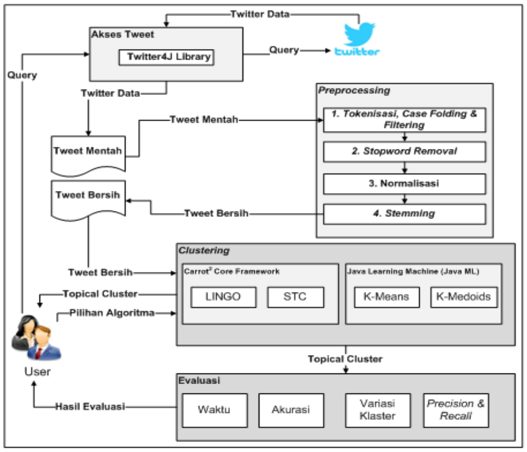

Fig. 1. System of architecture

Libraries which will be used in the process of developing the application can be seen in Figure 1. In the system of the architecture, there blocks that describe the processes that exist in each block the libraries used can be seen in the system. The processes that occur in the system are as follows:

a. Accessing Tweet

To access tweets from twitter in accordance with the needs of this research, it is necessary to have kay words with the theme Ambon tourism. As to the process of accessing tweet, libraries contained in Twitter4J are used.

b. Preprocessing

This process aims at converting raw tweets into clean tweets. Net Tweets is a tweet which can already be processed. The preprocessing stage is as follows:

- Tokenisasi

  This process is a process for breaking tweets into tokens.

- Case folding

  This process aims at equating character formats, for example all characters are coverted into lower case (smallletters) or upper case (capital letters). For example, given a tweet as shown in figure 2. After passing the case of case folding as shown in figure 3.

@zebipaskal | Fri Mar 23 22:41:15 ICT 2018 | rujak natsepa AMQ memang paling mantap… https://t.co/0h15oboRpE

Fig. 2. Tweet example

@zebipaskal | fri mar 23 22:41:15 ict 2018 | rujak natsepa amq memang paling mantap… https://t.co/0h15oborpe

Figure 2 Example of case folding stage results

a. Filtering

Filtering is useful for removing invalid characters (numeric and symbol characters). An example of this process can be seen in figure 3.

rujak natsepa amq memang paling mantap

Fig. 3. Sample of filtering step results

- Stopword removal

  It is a process of omission of words that do not contribute much to the contents of tweet. An example of a stop word removal step can be seen in figure 4.

rujak natsepa amq mantap

Fig. 4. Example of stopword removal step results

- Normalization

  The process of normalization in this research is a process to determine the words which are sinonim or has the same meaning and replaced with a particular word. For example, Ambonese people usually write the word Ambon as 'AMQ'. In this normalization process later, when the word 'AMQ' is found, it will be converted later into the word 'Ambon'. Example of this Stage, can be seen in figure 5.

rujak natsepa ambon mantap

Fig. 5. Example of normalization step results

- Stemming

  Stemming is a process of returning a word with its basic form. Stemming is a basic text processing tool that is often used to get an effective and efficient performance in text retrieval and text classification. An example of this stage result can be seen in figure 6.

rujak natsepa ambon mantap

Fig. 6. Example of stemming stage results

c. Clustering tweet

The Clustering Tweet process is used to determine the topics to be discussed in a set of tweets that have been accessed processed. The following are examples of the use of each algorithm in the process of tweet clustering, when the tweets which have been in the preprocessing stage are known, as shown on table 1.

Table 1: Example of a tweet for the clustering process

No	Contents of clean tweets
1	Pantai natsepa
2	Rujak natsepa
3	Makan rujak natsepa

Based on the tweet data shown on table 1, weighted matrix as shown on table 2 can be made.

Table 2: Tweet data on Weight Matrix

Dokumen/ Tweet	Pantai	Natsepa	Rujak	Makan
1	0.24	0	0	0
2	0	0	0.15	0
3	0	0	0.1	0.16

The clustering process for each algorithm is as follows:

a. Lingo

Based on the weighted matrix on table 2, the first step that should be done is to determine an abstract concept using the Singular Valve Decomposition oncept. In this way, the U – Matrix will be found as the following:

$$U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5476 & -0.8367 \\ 0 & 0.8367 & 0.5476 \end{bmatrix}$$

From the data of table 1, many phrases appear:

P1 = rujak natsepa

Using the next phrase P1 will create a weighted P matrix based on the phrase P1 as follows:

$$P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0.15 \\ 0 & 0 & 0.1 \end{bmatrix}$$

In the next step. Matrix M will be formed, with size a x b in which a indicates the number of phrases, b the number of tweets, Therefore the value a =1 and b = 3, whereas the value of M is determined by summing up the entire distance of each column on matrix U and P. The distances are calculated by using equation 3, so that Matrix M is obtained as follows:

$$M = \begin{bmatrix} 0 & 0 & \mathbf{1.19} \end{bmatrix}$$

Matrix P shows that the tweet containing the phrase ‘rujak natsepa’consists of 2 tweets, which are tweet 2 and tweet 3. Using equation 1, the cluster score = 1.19 x 2 = 2.38. In the Lingo algorithm, documents or tweets that are not included in any topic will be included in the “other topic”cluster.

b. STC

Based on the data on Table 1, the base cluster which is formed is like those shown on table 3. Table 4 shows the calculation of scores for each base cluster.

Table 3: Examples of STC Base Clusters

Base cluster	Phrase	Appear on tweet
A	Rujak natsepa	2,3
B	Natsepa	1,2,3

Table 4: Example of Calculation score base cluster

Result of Word/Phrase	**|B|** (total of document)	**f|P|** (number of words of user’s phrases)	**s(B)** (score cluster)
Rujak natsepa	2	2	4
Natsepa	3	1	3

In this way, two clusters are formed, namely ‘rujak natsepa’ and ‘natsepa’. Members of the cluster ‘rujak natsepa’ are tweet 2 and tweet 3, whereas members of the cluster ‘natsepa’ are tweets 1, 2, and 3.

c. K-Means

Supposing the specified number of cluster is 2 with the initial clusters are tweet 1 and tweet 2, supposing also, the terms symbolized as : pantai = U, natsepa = V, rujak = X, and makan = Y, then when using the formula Eucliedian Distance, the process of calculating the distance between tweet 1 and tweet 3, for example, will be done:

$$\sqrt{(U_1 - U_3)^2 + (V_1 - V_3)^2 + (X_1 - X_3)^2 + (Y_1 - Y_3)^2}$$

$$= \sqrt{(0.24)^2 + 0 + (-0.09)^2 + (-0.06)^2} = 0.63$$

The distance between the tweets generated is shown on table 5

Table 5: The distance between the tweets

Tweet	Tweet 1	Tweet 2	Minimum distance	Movement
1	0.24	0.58	0.24	*Cluster* 1
2	0.58	0.3	0,3	*Cluster* 2
3	0.63	0,25	0,25	*Cluster* 2

This, the resulting 2 clusters, cluster 1with members tweet 1 And cluster 2 with members tweet 2 and 3.

d. K-Medoids

Using the data on table 2, 2 cluster centers, for example, tweet 1 and tweet 3, will be initialized. When assumed that tweets 1 and 3 are medoids, so that C1 = (0,24, 0, 0, 0)and C2 = (0, 0, 0.1. 0.16) the object distance of the non-medoids towards the Cluster Center will be computed, using the distance formula Eucledian Distance on formula (3). The result of the computation of Distance non-medoids to medoids can be seen on table 6.

Tabel 6: Computation of Distance 1

Tweet	Distance	Cost
C1		
1	**0.24**	**0**
2	0.28	0.04
3	0.26	0.02
C2		
3	**0.26**	**0**
1	0.26	0
2	0.05	0.21
Total Cost C1		0.06
Total Cost C2		0.21

The next step is to choose non medoids from tweets other than tweet 1 and tweet 3. Since there is only Tweet 2 which is left, Tweet 2 is chosen.

Table 7: Calculation of distance 2

Tweet	Distance	Cost
Non Medoids		
2	**0.15**	**0**
1	0.28	0.07
3	0.05	0.1
Total Cost C2		0.17

When the total difference between non-medoids and medoids is below 0, the process stops. As the non-medoids only amounts to 1 tweet, then of medoids with the smallest cost is tweet 1 as C1 and Tweet 2 as C2. From this process, clusters with each of their members are obtained:

C1 : tweet 1
C2 : tweet 2 dan tweet 3.

4. RESEARCH RESULT

In the following sections a testing process will be conducted, and the results obtained in the testing process will be discussed. The test will include a number of aspects, namely processing time, precision, purity, recall and accuracy.

In this test, the number of tweets to be processed was 1000 tweets, consisting of 390 tweets with beach tour themes, 215 tweets with culture tour themes, 325 tweets with culinary themes and 70 other tourist themes.(Example: Jembatan Merah Putih, Gong Perdamaian etc.). The result of the test based on the time aspect can be seen on table 7.

Table 7: Test results based on processing time

Try Out to	Time (Mili Second)			
	Lingo	STC	K-Means	K-Medoids
1	1638	166	577	259
2	1643	89	595	87
3	1822	56	561	133
4	1520	64	561	103
5	1607	73	619	87
6	1672	122	593	87
7	1534	98	564	71
8	1583	97	579	87
9	1564	56	561	103
10	1578	73	562	91
Average	1616.1	89.4	577.2	110.8

The test results on the aspect of processing time or performance shows that the STC algorithm is the fastest algorithm compared to other algorithms. The STC algorithm needs an average of 89,4 mili seconds to cluster 1000 tweets. Whereas the Lingo algorithm has the longest average processing time, which is 1616.1 mili seconds to cluster 1000 tweets.

The following is the test mechanism for aspects other than time:

a. Purity

Purityis used to determine the quality of clustering results. Purityity indicates that all of the cluster members are members of a dominant class or a class with the most number of members in the cluster. The purity value is obtained by summing up the number of the dominant cluster members or the majority of the whole cluster which is later divided by the number of the document (Manning et al, 2008) or when formulated will be:

$$\text{Purity} = \frac{1}{N}\sum_{i=1}^{k} max_i \left|w_i \cap c_j\right| \qquad (1)$$

In which:

N : number of documents

k : number of cluster. The number of clusters for Lingo and STC algorithms are developed using the Carrot 2 Framework which is automatically determined by the algorithm, while the K-Means and K-Medoid algorithms are determined by the user.

w_i : *cluster* ke-i

c_j : the set of documents in the cluster c_j

b. Precision, recall and Accuracy

When the test data of 100 tweets and the topic for each tweet are known, that condition can be referred to as the actual condition. While the results obtained through the clustering process in which each document would be inserted into certain clusters, the condition can be said as predictive conditions or predictive value.

Table 8 is a contingency matrix that describes the possibilities that arise when a comparison between actual conditions and predictive condition is made.

Tabel 8: Contigency Matrix

		Actual Condition	
		True	False
Condition Prediction	True	True Positif (TP)	True Negatif (TN)
	False	False Positif (FP)	False Negatif (FN)

Related to the Clustering results, then:

TP : The total number of tweets on any cluster that was generated and it is at the proper cluster corresponding to actual conditions.

TN : The total number of Tweets on any cluster that was generated but is not corresponding to actual conditions.

FP : The total number of Tweets which do not fit in any cluster on predictive condition but enter into a certain cluster in actual condition.

FN : The Total number of Tweets which do not fit into neither actual as well as predictive condition.

In this way, when formulated into:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2)$$

$$\text{Recall} \quad = \frac{\text{TP}}{\text{TP+TN}} \qquad (3)$$

$$\text{Accuracy} \quad = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$
(4)

The testing process with the system was implemented in testing Java class by applying the formula 8 to formula 8. Summary of the test results conducted can be seen on table 9.

Table 9: The result of Purity, Precision &Recall Testing and Accuracy

Algorithm	Purity	Precision	Recall	Accuracy	Average
Lingo	1	0.611	0.611	0.829	0.76275
STC	1	0.513	0.513	0.701	0.673
K-Means	0.551	0.102	0.102	0.382	0.28425
K-Medoids	0.461	0.092	0.092	0.312	0.23925

Testing on purity aspect shows that Lingo and STC algorithms are algorithms which have the best purity value, which is 1.Purity for Lingo and STC algorithms have value 1 which means that almost all of the documents or tweet are successfully placed in the appropriate cluster or in its dominant class. In testing the precision aspect, the Lingo algorithm succeeds in clustering the tweet precisely in the cluster corresponding to the actual condition.

In the recall aspect, the Lingo algorithm is an algorithm which give the best test result which is 0. 611. It means that the Lingo algorithm succeed other than to place tweets in clusters that match the actual conditions, could also minimized the number of tweet replacement error in clusters that do not match the actual conditions.

In the recall aspect, Lingo algorithm is an algorithm with the best accuracy, which is. STC algorithm also 0.829 also gives a fairly good accuracy, which is 0.701 and only 0.128 different from the Lingo algorithm. This indicates that the Lingo and STC algorithms are algorithms which could cluster to the maximum tweet into the test data.

5. DISCUSSION

In this section, the test results which have been done in the previous section will be discussed. Based on the time aspect, the test results can also be depicted into a line graph which can be seen on figure 7. From the test result of 10 experiments, it can be concluded that the Lingo algorithm takes a longer processing time compared to other algorithms with an average processing time of

1616.1 mili seconds .While the STC algorithm requires the smallest processing time with an average time of 89.4 mili seconds to process 1000Tweets.

Although the Lingo algorithm takes the longest time to process tweets, when viewed from the aspects of purity, precision, recall and accuracy, the Lingo algorithm gives the best result the result obtained by the Lingo and STC algorithms show that clustering algorithms that group documents based on phrase similarity will provide better result. In clustering text documents such as tweets. It can be seen that from the purity aspect, Lingo and STC algorithms get very good results, which is 1 and accuracy value above 90%.

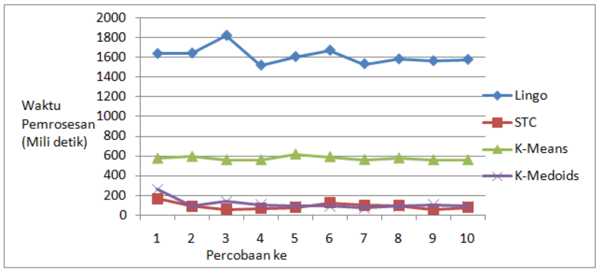

Fig. 7. Graph of the result of testing based on processing time

In addition to the above aspects, when viewed from the aspect of the number of clusters, the Lingo and STC algorithms cab generate a more flexible number of clusters depending on the number of processed text documents, while the K-Means and the K-Medoids algorithms the number of clusters are determined by the user.

Based on the results of testing and discussion above, it can generally be said that the algorithm STC as the best algorithm in clustering tweet-themed Ambon Tourism Tweet.

6. CONCLUSION

Based on the results of testing and discussion it can be concluded that:

a. Lingo algorithm is an algorithm which takes the longest time in the process of clustering with the average processing time of 1616.1 to process 100 tweets, while STC is the fastest algorithm with the processing time of an average of 89.4 mili seconds.
b. Based on the testing process using 100 data tweet, that the topics have been determined, it can be concluded that based on the purity and accuracy aspects, Lingo algorithm gives the best result in the purity value of 1 and the accuracy value of 0.829. The result obtained by STC algorithm can be said very well, with the purity value of 1and the accuracy of 0.701.
c. Clustering algorithm is to group documents based on similarity of phrases,, Like Lingo and STC algorithms. Clustering algorithm gives better results in grouping text documents.
d. To group text documents, it is advisable to use Lingo and STC Algorithms, since they can give better results, and they can group documents to a more flexible number of clusters, depending on the number of documents to be grouped.

Research. This study recommends Suffix Tree Clustering (STC) algorithm to be used in determining the topics of text documents in the form of a tweet with the theme Tour of Ambon.

REFERENCES

[1] Andonovski & Chorbev (2014) in a study titled “Clustering of tweets from popular athletes”: implement an algorithm to define the most popular topics of famous athletes. ICT Innovations 2014 Web Proceedings ISSN 1857-7288.

[2] Ashktorab, Z., Brown, C., Nandi, M. & Cullota, A., 2014, Tweetdr: Mining Twitter to Inform Disaster Response, Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA, May 2014.

[3] Saravanou, A. & Valkanas, G., 2015, Twitter FLoods when if Rains: A Case Study of the UK Fllods in early 2014, http://cgi.di.uoa.gr/~gvalk/pubs/TwitterFloods.pdf, diakses tanggal 2 Agustus 2017.

[4] Sechelea, A. Huu, T.D., Zimos, E. & Deligiannis, N., 2015, Twitter Data CLustering and Visualization, http://homepages.vub.ac.be/~ndeligia/pubs/TwitterDataClusteringVisual.pdf, diakses tanggal 1 Agustus 2017.

[5] Godfrey, D., Johns, C., Sadek, C., Meyer, C. & Race, S., 2015, A Case Study in Text Mining: Interpreting Twitter Data From Worlds Cup Tweets, http://meyer.math.ncsu.edu/Meyer/PS_Files/CaseStudyInTextMining.pdf, diakses tanggal 2 Agustus 2017.

[6] Friedemann, V., 2015, Clustering a Customer Base using Twitter Data, http://cs229.stanford.edu/proj2015/310_report.pdf, diakses tanggal 1 Agustus 2017.

[7] Unnisa, M., Ameen, A., & Raziuddin, S., 2016, Opinion Mining on Twitter Data using Unsupervised Learning Technique, International Journal of computer Applications, Vol. 148, No.12, August 2016.

[8] Alnajran,N., Crockett, K., McLean, D. & Lathan, A., 2017, CLuster Analysis of Twitter Data: A Review of Algorithms, https://e-space.mmu.ac.uk/617901/1/ICAART_2017_110_CR_final.pdf, diakses tanggal 01 Agustus 2017.