Application of Density Based Clustering of Disaster Location in Realtime Social Media

Mochammad Haldi Widianto [1], Ivan Diryana Sudirman [2], Muhammad Hanif Awaluddin [3]

[1]*Informatics Departement, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia*
[2] *Enterpreneurship Departement, BINUS Business School Undergraduate Program, Bina Nusantara University, Jakarta, Indonesia*
[3] *CV. Izz Digital Indonesia, Bandung, Indonesia*

Abstract **– Online life is used as a method of finding information, one of which is Twitter as the medium. The occurrence of natural disasters is very detrimental. Therefore, the application is needed to see natural disasters through social media Twitter. A small number of studies using clustering methods based on Twitter user data density are the beginning of this research. With the availability of data in certain areas makes it easy to group. After that, the data is grouped based on a high degree of similarity. One result of applying this method is the location of the disaster. NER-based rules are used to discover out the area of the disaster. Data accuracy testing is performed using the Silhouette coefficient.**

Keywords **– Density-based clustering, NER rule-based, location disaster, Silhouette coefficient.**

1. Introduction

Almost all people have social media accounts, which are used to share information. Twitter's social media is a highlight of information as it focuses more on text information and news information.

DOI: 10.18421/TEM92-13
https://doi.org/10.18421/TEM92-13

Corresponding author: Mochammad Haldi Widianto, *Informatics Departement, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia.*
Email: mochamad.widianto@binus.ac.id

Received: 24 December 2019.
Revised: 09 June 2020.
Accepted: 16 June 2020.
Published: 28 August 2020.

Tweets that report such events are usually covered by flooding a meaningless tweet, the algorithm should be measurable given the number of tweets there [1], [2]. Twitter itself comes from a tweet word that has the meaning of tweets. This is what underpins its users to do tweets (text elements) through social media Twitter. Tweet copy often contains current information, but not necessarily the actual data. Thus, an implementation is required that can parse real (true-to-occurring) tweet data and hoaxes (not occurring).

Indonesia is located between 3 tectonic plates [3], [4] the Eurasian plate, the Pacific Plate and the Indian-Australian Plate. This condition resulted in Indonesia being susceptible to natural disasters. The impact of natural disasters is detrimental to both the country and the community because natural disasters can damage the facilities and threaten human psyche. Therefore, in the event of a disaster, it should be quickly and accurately addressed. Social media Twitter can be used as a medium to know the natural disasters that are happening or even have occurred but have not been addressed. In previous studies [5], density-based clustering methods had connectivity and density capabilities to handle large datasets. Both in terms of accuracy, the memory utilization and quality are better than the DBSCAN method. The results of grouping use density-based clustering performed identification (locating and categorizing) entities in the given text that belong to a specified category or class [6]. It is called the NER rule-based, in which the study will be used to identify the location of natural disasters

2. Preliminary Analysis

2.1. Analysis of Problems

The initial stage is to analyze the problem. Data obtained from social media are generally realtime, but this data is not necessarily reliable and useful. The amount of data talking about disasters can be

both profit and loss. The advantage is for people who respond to a tweet copy about disasters by first analyzing the truth. It becomes a disadvantage when news about disasters are not explained in advance of the fact. The density-based clustering method is the solution to solve the problem, and it is better than other clustering methods. Therefore, a density-based clustering method is very suitable for grouping data that has very close data similarity. Results grouping data can be utilized to analyze the location of events. The purpose of the event location analysis is to ensure precisely that the disaster is happening.

Therefore a critical point can be taken if the problem that occurs from the description is that data about a disaster is very beneficial to the community if the distribution of data can be grouped using applications with excellent and correct like a density-based clustering method.

2.2. Research Methods

The application of the density-based clustering method will generate a value of grouping result quality using a density-based clustering method. Because it produces a profit from a variety of grouping results, this research includes a quantitative process. The research steps can be seen in Figure 1.

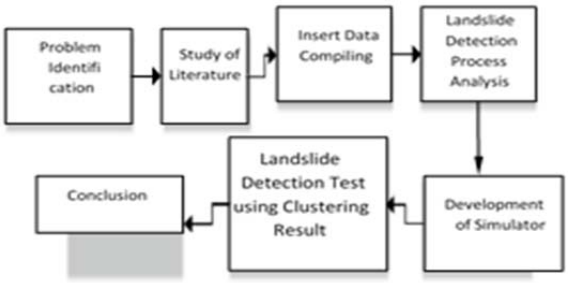

Figure 1. Research methods

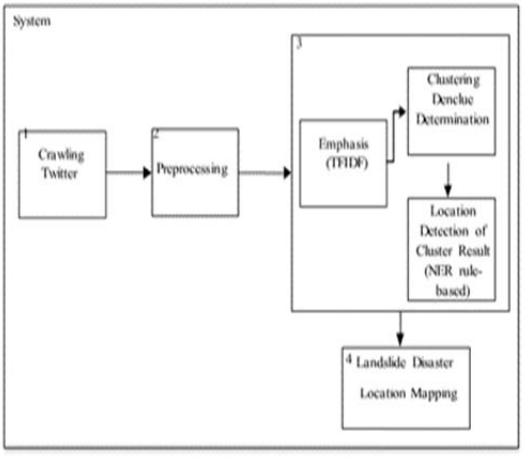

Figure 2. Stages of data analysis

After analyzing the data, the next step is to do the simulator design. The simulator design is done based on an agile model [7], [8]. Here Figure 3 is the application of agile models on this research.

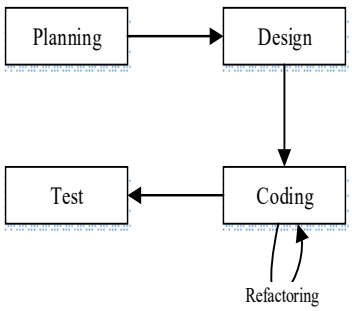

Figure 3. Model agile

Based on Figure 3, the initial stage is planning against the simulator design. After planning, the next stage is the design or mockups that will be used as a simulator design. After doing the design or mockups, the next step is to do the coding based on the programming language used, and coding adjusts from the planning as well as the design/mockups. The coding phases are continuously improved, so it will be appropriate based on the planning and design/mockups. After coding, the next step is to test the coding results. Testing aims to assess the outcome of applying density-based clustering methods. After trial, the stage is reviewing the results of the landslide location.

3. Algorithm Needs Analysis

3.1. Crawling Data Twitter

The Twitter Developer Platform provides many products, tools, and API resources that allow you to leverage the power of Twitter's, global, and real-time Twitter communications networks [9]. The following figure 4 is a flow analysis diagram of the collection phase.

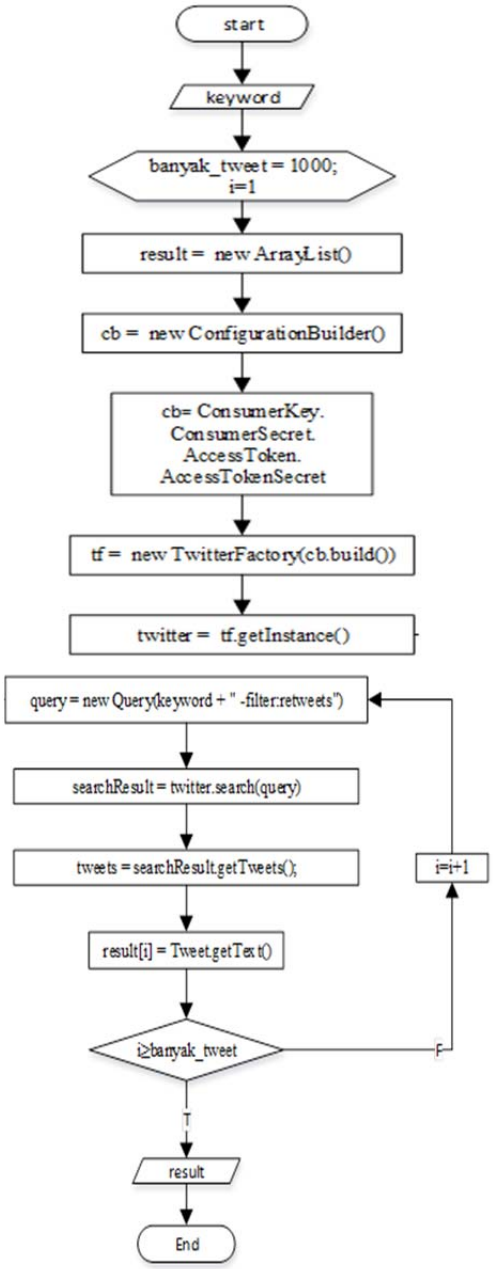

Figure 4. Flowchart Twitter data collection

The explanation of Figure 4 is:

1. Enter or input the destination keyword
2. Initializing the amount of tweet data with up to 1000 data.
3. The process of creating new objects is inserted into the result variable.
4. Object creation to authenticate with Twitter.
5. Inserting the Twitter Key API.
6. The initialization creation Twitter object linked via CB variable.
7. Process for lowering classes available on the Twitter API
8. The process for creating a query against the keyword, where the search keyword request does not contain retweets.
9. Searching Twitter data query requested.
10. Search Twitter data by reading user status.
11. Storing tweet data.
12. Condition whether I ≥ banyak_tweet, if it is accurate and performs step 13 (state true) and if it is false, then do step 8 (condition false).
13. Output results are Twitter crawl data sent in the result variable in the form <string>

The following Table 1. is the Twitter test data conducted in this study.

Table 1. Test Data

Keyword	Test Data
"landslide"	The condition of a house buried by a landslide in the 10-metre-high Cliffs of South Tangerang http://ht.ly/CncC30g0qx9
	Floods and Landslides in Jembrana, hundreds of Residents Isolated offices ... https://t.co/5LnMEhbk1x #jembrana #balitoday
	The condition of a house buried by a landslide in the 10-metre-high Cliffs of South Tangerang
	Landslide in Jembrana, hundreds of residents of Isolated #longsor
	this landslide of creepy in Tangerang :(

3.2. Preprocessing

Researchers took a case by taking the main word for a landslide. Here Figure 5 is one example of a tweet copy of a natural disaster.

Figure 5. Twitter Data containing natural disasters

Preliminary data taken about the disaster is still not optimal because it still has elements that will not affect the process of grouping data. Examples of data that will not take effect when the process grouping data is the use of URL addresses, use of hashtags and use of tagging or mentions used in tweet data. Here Figure 6 is an example of a tweet containing the use of URLs, hashtags and mentions.

Figure 6. Twitter Data that has elements of using URLs, hashtags and mentions

Tweet data can also have punctuation elements, and even these punctuation marks often have a meaning of the expression. Examples of tweet data with punctuation elements can be found in Figure 7.

Figure 7. Twitter Data with punctuation elements

The preprocessing process is used as a data modifier to a format that matches the purpose of being a proper form. There are several stages of preprocessing that include case folding, cleaning, tokenizing and filtering. The case folding is here for the conversion of capital letters (uppercase) to lowercase the intention of standardising the word. "LowerCase ()" is the function of the folding case to lowercase. Cleaning is here to clean some components. There are several components of the tweet data to be cleaned or deleted, i.e. "@", Link, "#" and punctuation. "Cleaning ()" is a function for cleaning those components.

The following Table 2. is the result of the preprocessing of the test data.

Table 2. Result of preprocessing

No.	Data *Preprocessing*
d1	kondisi rumah tertimbun longsor tebing setinggi meter tangerang selatan
d2	banjir longsor jembrana ratusan warga terisolasi kantor
d3	kondisi rumah tertimbun longsor tebing setinggi meter tangerang selatan
d4	longsor jembrana ratusan warga terisolasi
d5	longsor tangerang menyeramkan

3.3. TF-IDF

TF-IDF is included in the process of counting the number of words. The word appears in each document [10], [11]. Equation 1 is no TF-IDF formula.

$$TFIDF\left(d_{i_{t_h}}\right) = TF\left(d_{i_{t_h}}\right) \times \log\frac{N}{DF(t_h)} \quad (1)$$

Descriptions :

d_i = document to-i
i = number of documents
h = number of words in the whole document
t_h = term to-h
$TF\left(d_{i_{t_h}}\right)$ = number of occurrences t_h on d_i
log = logarithm natural base 10
N = amount of documents
$DF(t_h)$ = number of documents that have a term t_h
$TFIDF\left(d_{i_{t_h}}\right)$ = the frequency of t_h occurrence in the document d_i

The following Table 3. is the result of the TF-IDF implementation of document 1.

Table 3. Result Implementations TF-IDF

$\mathbf{t_h}$	Term	$\mathbf{DF(t_h)}$	$\boldsymbol{x_1 = WD\left(d_{1_{t_h}}\right)}$
t_1	banjir	1	0
t_2	jembrana	2	0
t_3	kantor	1	0
t_4	kondisi	2	0.397940009
t_5	longsor	5	0
t_6	menyeramkan	1	0
t_7	meter	2	0.397940009
t_8	ratusan	2	0
t_9	rumah	2	0.397940009
t_{10}	selatan	2	0.397940009
t_{11}	setinggi	2	0.397940009
t_{12}	tangerang	3	0.22184875
t_{13}	tebing	2	0.397940009
t_{14}	terisolasi	2	0
t_{15}	tertimbun	2	0.397940009
t_{16}	warga	2	0

3.4. Density-Based Clustering

3.4.1. Preclustering

The first process of the density-based grouping method follows the procedure [12]. This process has the purpose of creating a folder. For each data point, with the function as the first step towards the calculation of the density function. Step before clustering consist of 3 steps, namely:

a) Construct the Hyper Rectangle

In the development of hyper rectangle, the value of σ (sigma) and ξ (minPts) is required. The value used in this study is σ = 0.5 and ξ = 1. Calculations are done using the B +-tree function in order to find the

value of the pyramid on each word with 2σ length. Equation 2 is the pyramid value formula.

$$pv(x_{j_{t_h}}) = (i + 0.5 - x_{j_{t_h}} \; mod \; 2\sigma)(2)$$

The following Table 4. is the search results pyramid value t_1 against the whole document.

Table 4. Pyramid value Search results t_1 against all documents

Key B+tree	
x_i	t_1
x_1	1.5
x_2	0.8
x_3	1.5
x_4	1.5
x_5	1.5

Pyramid value Search is done on the whole word against each document, thus generating a length is 16, and the height is 5.

b) Determine the populated Cubes

Each population in the cube will be given an index or principal value that is the result of the hyper-rectangle formation phase. Here Figure 8 is the use position of key and data points.

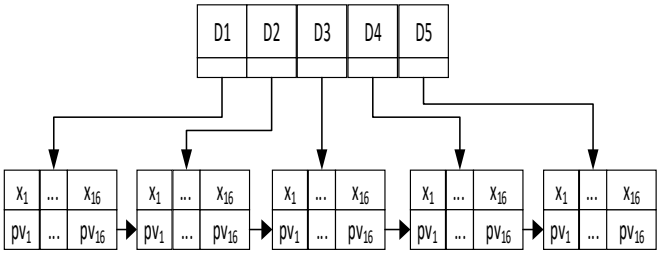

Figure 8. Key usage positions and data points

In the density-based clustering method, the implementation of the B + Tree function aims as data input to be more structured. The following Table 5. is the result of implementing the first word B+ tree function against the whole document.

Table 5. Results of implementation of B + tree function

	t_1
x_1	0
$pv(x_{1_{t_h}})$	1.5
x_2	0.69897
$pv(x_{2_{t_h}})$	0.8
x_3	0
$pv(x_{3_{t_h}})$	1.5
x_4	0
$pv(x_{4_{t_h}})$	1.5
x_5	0

The implementation of the B + Tree function is done in all words to the whole document.

c) Construct a map by connecting the populated cubes

Build a plan by linking the population; the initial stage is to find the distance of data on each data. Here is an example of x_1 distance calculation against x_2.

$$d(x_1, x_2) = \sqrt{(|x_{1_{t1}} - x_{2_{t1}}|^2 + .. + |x_{1_{t_h}} - x_{2_{t_h}}|^2)}$$

$$d(x_1, x_2) = \sqrt{(|0 - 0.698970004|^2 + \ldots . + |0 - 0.397940009|^2} = 1.663807007$$

The distance calculation is done against all data.

3.4.2. Clustering

After doing before clustering, the next step is to perform the clustering process. The clustering process is as follows:

a) Specifying A Density Value Attractors

The stages in determining the density attractors can be seen in Figure 9.

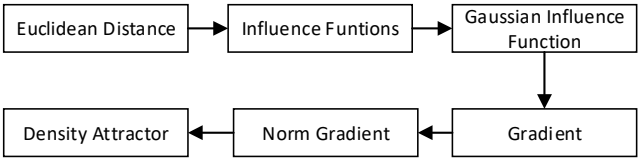

Figure 9. The process of determining density attractors

Here Equation 3 is the formula of Euclidean distance, Equation 4 is the formula of the influence function, equation 5 is the formula of Gaussian influence function, Equation 6 is the formula of gradient, Equation 7 is the formula of norm gradient and Equation 8 is the formula of density attractors.

$$d(x_j, x_i) = \sqrt{(x_{j_{t_1}} - x_{i_{t_1}})^2 + \cdots + (x_{j_{t_h}} - x_{i_{t_h}})^2} \quad (3)$$

$$f_{gauss}(x_j, x_i) = e^{-\frac{d(x_j, x_i)^2}{2\sigma^2}} \quad (4)$$

$$f^D_{Gauss(x_j)} = \sum_{i=1}^{N} e^{-\frac{d(x_j, x_i)^2}{2\sigma^2}} \quad (5)$$

$$\nabla f^{\;D}_{Gauss}(x_j) = \sum_{i=1}^{N} (x_i - x_j) * e^{-\frac{d(x_j, x_i)^2}{2\sigma^2}} \quad (6)$$

$$\left|\left|\nabla f^D_B(x_j)\right|\right| = \sqrt{|\nabla f^{\;D}_{Gauss}(x_j)|^2} \quad (7)$$

$$x = x_0;\ x_j = x_{j+1} + \delta \frac{\nabla f_B^D(x_{j+1})}{\left\| \nabla f_B^D(x_{j+1}) \right\|} \quad (8)$$

After the calculation of local density attractors, the next step is detecting the cluster pullers. If $f(x_{j+1}{}^*) > f(x_j{}^*)$ is correct the calculation is terminated and $f(x_{j+1}{}^*) = x^*$ (towing *cluster)*. Then x^{*0} act as a cluster that draws a local density value of 1.38520245.

b) Result Grouping Density-Based Clustering

Value x^* a cluster puller. Before performing cluster detection, the local density value of x Performed noise detection. If $f(x_j{}^*) \geq \xi$, then x^* not including data *noise*. Value *local density attractor* x = 1.38520245, dan value $f(x_j{}^*) \geq \xi$ then not including noise.

The following Table 6. is the result of grouping data using density-based clustering.

Table 6. Final Result Cluster

Data Tweet	$f^D_{Gauss(x_j{}^*)}$	Cluster
1	1.38520245	1
2	1.3520932	2
3	1.38520245	1
4	0.825619403	*Noise*
5	1.355664363	2

3.5. Silhoutte Coefficient

Silhouette coefficient is one of the measures in the Intristic method. Silhouette coefficient is also referred to as Silhouette index. Here is the formula of Sillhoutte coefficient [13],[14].

$$a_j^r = \frac{1}{C_r - 1} \sum_{i=1}^{C_r} d(x_j^r, x_i^r) \quad (9)$$

$$i = 1,2, \dots, c_r$$

$$b_j^r = min \left\{ \frac{1}{C_{\sim r}} \sum_{i=1}^{C_{\sim r}} d\left(x_j^r, x_i^{\sim r}\right) \right\}, i \quad (10)$$

$$= 1,2, \dots, C_{\sim r}$$

$$SC_j^r = \frac{b_j^r - a_j^r}{\max\{a_j^r, b_j^r\}} \quad (11)$$

$$SC = \frac{1}{k} \sum_{j=1}^{k} SC \quad (12)$$

The smaller the value a_j^r, the cluster becomes denser. Meanwhile, the value b_j^r describes how far the j object is with other objects from different clusters. So, if a_j^r Very small value and b_j^r very big value then Silhouette coefficient will approach 1. The following Table 7. is a measurement of the silhouette coefficient value.

Table 7. Size Of Silhouette Coefficient Value

Silhouette Coefficient	Interpretation
0.7 < SC <= 1	Strong clusters (strong structure)
0.5 < SC <= 0.7	The cluster has decent or appropriate (medium structure)
0.25 < SC <= 0.5	Weak clusters (weak structure)
SC <= 0.25	Cannot be said as a cluster (no structure)

The results for the first cluster are:

$$SC_j^1 = \frac{b_j^1 - a_j^1}{\max\{a_j^1, b_j^1\}}$$

$$SC_j^1 = \frac{0.631872776 - 0}{0.631872776} = 1$$

And the result for the second cluster is:

$$SC_j^2 = \frac{b_j^2 - a_j^2}{\max\{a_j^2, b_j^2\}}$$

$$SC_j^2 = \frac{0.631872776 - 2.931429051}{2.931429051} = -0.784448893$$

Once the SC values are obtained from all the clusters, the overall SC results are based on the equation (2.15):

$$SC = \frac{1}{k} \sum_{j=1}^{k} SC$$

$$SC = \frac{1}{2} (1 + (-0.784448893))$$

$$SC = \frac{1}{2} (0.215551107)$$

$$SC = 0.1077755535$$

The results of the coefficient of the 0.1077755535 were held against 5 test data and included no structure.

3.6. NER Rule-Based

NER rule-based [15], [16] is an identification that has a base to the rules with a function in order to determine the location. The rules used in this study can be seen in Table 8.

Table 8. Words Used As A Rule

Rules Used	
di	kabupaten
pada	Kab
negara	Provinsi
jalan	Prov
kota	Lokasi

Based on Table 8., the word rule will be used to detect the location of the landslide event in nine words. The following Table 9. is the result of location detection against cluster results.

Table 9**.** Location Detection Of Cluster Results

Cluster	Tweet	Locations
1	**Kondisi Rumah Tertimbun Longsor Tebing setinggi 10 Meter di Tangerang** Selatan http://ht.ly/CncC30g0qx9	[**Tangerng**]
1	**Kondisi Rumah Tertimbun Longsor Tebing setinggi 10 Meter di Tangerang** Selatan	
2	Banjir dan Longsor di Jembrana, Ratusan Warga Terisolasi – Kantor… https://t.co/5LnMEhbk1x #jembrana #balitoday	[**tangerang,Jembrana**]
2	ini longsor di 935cenario935 menyeramkan☹	

Catastrophic events are identified by 2 occurrences as well as 2 locations.

4. Result

The implementation stage is done using the sigma values of 0.25, 0.5, 0.75, 1 and minPts, namely 1, 2, 3. Here Figure 10 is the result of designing an application from this research.

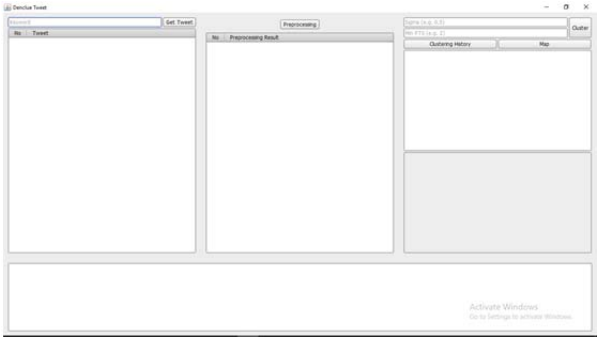

Figure 20. Application design results

The first implementation was done in real-time on 01 November 2019. The following Table 10. is the result of implementation of Twitter data, where the data obtained is as many as 100 tweets containing the keyword “banjir”.

Table 10. Implementating keyword “banjir”

No	σ	ξ	Amount Noise	Amount Cluster	Quality Cluster	Locations
1	0.25	1	82	4	*weak structure*	2
2		2	28	2	*Strong structure*	1
3		3	28	2	*Strong structure*	1
4	0.5	1	68	3	*no structure*	1
5		2	67	3	*no structure*	1
6		3	67	3	*no structure*	1
7	0.75	1	45	4	*No structure*	2
8		2	47	3	*No structure*	2
9		3	51	3	*No structure*	2
10	1	1	25	5	*No structure*	2
11		2	22	2	*No structure*	2
12		3	18	*2*	*No structure*	2

Based on Table 11., it is known that the best implementation is in 2nd and 3rd implementations.

The following Table 11. is the result of the implementation of Twitter data, where the data obtained in as many as 94 tweets containing the keyword “gempa bumi”.

Table 11. Implementing keyword “gempa bumi”

No	σ	ξ	Amount Noise	Amount Cluster	Quality Cluster	Locations
1	0.25	1	70	10	*strong structure*	4
2		2	58	2	*strong structure*	1
3		3	-	-	-	-
4	0.5	1	57	13	*medium structure*	4
5		2	73	5	*strong structure*	3
6		3	82	2	*strong structure*	1
7	0.75	1	42	13	*weak structure*	4
8		2	55	8	*medium structure*	3
9		3	64	4	*No structure*	3
10	1	1	18	10	*No structure*	4
11		2	27	4	*No structure*	2
12		3	33	2	*No structure*	2

Targeting Table 11., known to the best implementation is in the 1st, 2nd, 5th, 6th implementation of the results.

Based on Table 10. and Table 11., it can be concluded that the smaller value of the σ parameter, the closer the data similarity. And the greater the value of ξ, the more data on a group.

5. Conclusion

Based on the results of the analysis, and application of density-based clustering method to the location of disaster on social media in realtime. It can be concluded that the use of density-based clustering. This method is beneficial to classify data based on its resemblance level, and NER rule-based can detect the results of each tweet already grouped using a density-based clustering method. The smaller the σ parameter value (distance), the closer the data similarity. And the higher the amount of ξ (minimum member), the more data in a group. The rule-based NER relies heavily on the specified word rule to classify a location.

References

[1] Jiang, L., Shi, L., Liu, L., Yao, J., Yuan, B., & Zheng, Y. (2019). An efficient evolutionary user interest community discovery model in dynamic social networks for Internet of people. *IEEE Internet of Things Journal*, *6*(6), 9226-9236.

[2] Xiaolong, D. E. N. G., Jiayu, Z. H. A. I., & Luanyu, Y. (2017). Vector Influence Clustering Coefficient Based Efficient Directed Community Detection Algorithm. *Journal of Electronics & Information Technology*, *39*(9), 2071-2080.

[3] Tantri, A. H., & Rakhmawati, N. A. (2019, July). Designing A Natural Disaster Ontology for Indonesia. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)* (pp. 130-134). IEEE.

[4] Puspita, I. A., Soesanto, R. P., & Muhammad, F. (2019, April). Designing Mobile Geographic Information System for Disaster Management by Utilizing Wisdom of The Crowd. In *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)* (pp. 496-500). IEEE.

[5] Ramesh, D., & Kumari, K. (2018, March). DEBC-GM: denclue based gaussian mixture approach for big data clustering. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-8). IEEE

[6] Xiaowei, W., Longbin, J., & Jialin, M. (2008, November). Use of NER Information for Improved Topic Tracking. In *2008 Eighth International Conference on Intelligent Systems Design and Applications* (Vol. 3, pp. 165-170). IEEE.

[7] Boehm, B., Rosenberg, D., & Siegel, N. (2019, July). Critical Quality Factors for Rapid, Scalable, Agile Development. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (pp. 514-515). IEEE.

[8] Chiyangwa, T. B., & Mnkandla, E. (2018). Agile methodology perceived success and its use: The moderating effect of perceived compatibility. *South African Computer Journal*, *30*(2), 1-16.

[9] Z. Doshi, S. Nadkarni, K. Ajmera, and N. Shah. (2018). TweerAnalyzer: Twitter Trend Detection and Visualization, *2017 Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2017*.

[10] Yamout, F., & Lakkis, R. (2018, September). Improved TFIDF weighting techniques in document Retrieval. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)* (pp. 69-73). IEEE.

[11] Suzuki, Y., Mitsukawa, M., & Kawagoe, K. (2008, September). A image retrieval method using TFIDF based weighting scheme. In *2008 19th International Workshop on Database and Expert Systems Applications* (pp. 112-116). IEEE.

[12] He, J., & Pan, W. (2010, March). A Denclue based approach to neuro-fuzzy system modeling. In *2010 2nd International Conference on Advanced Computer Control* (Vol. 4, pp. 42-46). IEEE.

[13] Zoubi, M. D. B. A., & Rawi, M. A. (2008). An efficient approach for computing silhouette coefficients. *Journal of computer science*, *4*(3), 252.

[14] H. Řezanková. (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. *21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics 2018* Kutná Hora, Czech Republic 29 August 2018 – 2 September, pp. 1-10.

[15] Kejriwal, M. (2019). Information Extraction. In *Domain-Specific Knowledge Graph Construction* (pp. 9-31). Springer, Cham.

[16] Ferreira, J., Oliveira, H. G., & Rodrigues, R. (2019). NLPyPort: Named Entity Recognition with CRF and Rule-Based Relation Extraction. In *IberLEF@ SEPLN* (pp. 468-477).