Journal of Physics:
Conference Series

PURPOSE-LED
PUBLISHING™

PAPER • **OPEN ACCESS**

Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms

To cite this article: Mustakim *et al* 2021 *J. Phys.: Conf. Ser.* **1783** 012016

View the article online for updates and enhancements.

Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms

Mustakim[1*], Muhammad Zakiy Fauzi[2], Mustafa[3], Assyari Abdullah[4], Rohayati[5]

[1,2] Departement of Information System, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
[3,4,5] Departement of Communication, Faculty Dakwah and Communication, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
[1,2]Puzzle Research Data Technology, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia

*mustakim@uin-suska.ac.id

Abstract. Natural disasters are disasters caused by events or series of events caused by nature such as earthquakes, tsunamis, volcanic eruptions, floods, tornadoes, and landslides. Some of these natural disasters have taken a lot of public attention, from empathy, sadness and criticism that form an opinion on social media. One of the most popular social media used by the public is Twitter. Opinions written by Twitter users are called tweets. A collection of tweets can be processed to obtain information by using data mining techniques namely Text Mining. In this study, the Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm and K-Medoids were used. The result of this study shows that DBSCAN is the best algorithm because it has the Silhouette Index (SI) validity of 0.9140 and the average execution time in RapidMiner Studio is 83.40 seconds. Meanwhile, the K-Medoids algorithm has a Silhouette Index (SI) validity of 0.2259 and an average execution time in RapidMiner Studio 849.93 seconds. The frequency of the word "earthquake" dominates for the positive category, the word "disaster" dominates the negative category, and the word "flood and earthquake" dominates the negative category.

1. Introduction

The use of social media applications among the people indicates that technology has developed very rapidly and has given great influence in terms of social interaction. With the use of social media, the interaction happens as if it has been moved into a virtual platform [1]. Some social media applications that are widely used by Indonesian people based on a survey from the Ministry of Communication and Information, namely Facebook with 65 million users, Twitter with 19.5 million users, Google+ with 3.4 million users, LinkedIn with 1 million users, and Path with 700 thousand users [2].

One social media that provides free API access is Twitter. Twitter is a social media application created in March 2006 by Jack Dorsey that gives users access to real-time information [3]. Active Twitter users currently reach 22% of internet users in the world with 500 million tweets per day that are dominated by 80% mobile devices [4].

With the availability of information access in real-time, Twitter users can share information about an event or certain events quickly, including the event ofnatural disasters [5]. Natural disasters are events that are caused by nature and are destructive such as floods, earthquakes, tsunamis, landslides, fires, and whirlwinds [6]. Indonesia is a country with a high intensity of natural disasters, this is caused by the Indonesia's geographical location which is in the confluence of three major plates of the world, namely the Eurasian, Pacific and Indo Australian plates [7]. Natural disasters trigger people to provide information and opinions in the form of tweets. These tweets can be in the form of empathy or complaints. Collection of tweets data can be processed by using one of the data mining techniques, namely text mining [8][9].

Text Mining is a discipline that studies text data processing such as information retrieval, document text analysis, data extraction, and data visualization [10]. Text Mining has a tendency in the field of data mining research, so there is similarity in architectural aspects [11]. One of the data mining techniques that are commonly used in text mining research is clustering. Clustering is a technique used to group data in a cluster (groups) using certain parameters so that object in one cluster has the same level of similarity [12]. The Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm and K-Medoid were used in this research.

Previous research on Analysis and Implementation of Community Detection Using the DBSCAN Algorithm on Twitter found that the advantages of the DBSCAN algorithm were to produce clusters that were able to handle noise / outliers, more accurate clusters results, and it was good for large amounts of data [13]. Other research on the K-Medoids algorithm is the Comparison of K-Means and K-Medoid Clusters on Outlier Data, with the results that the K-Medoids algorithm was better than K-Means in clustering data with 5% outliers.

2. Materials and Method

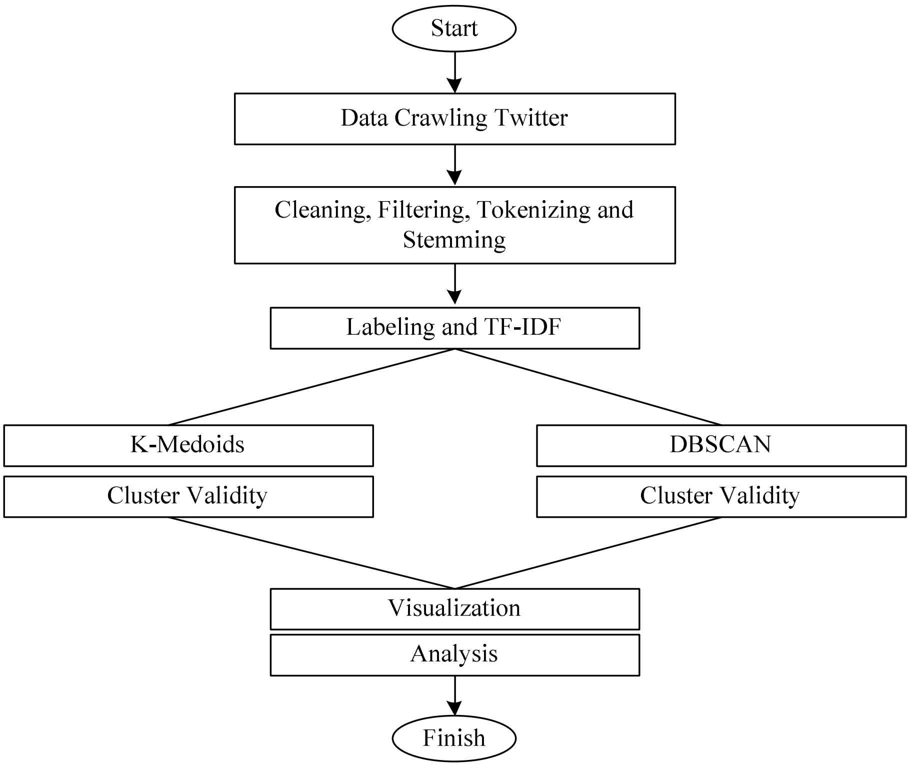

Figure 1. Research Method

The planning stage is the stages that must be carried out by researchers in conducting a study which consists of determining research objectives, identifying problems, determining the limitations of the study, reviewing literatures, and determining the data. The next stage is collecting data by crawling from Twitter and reviewing literatures. At the stage of preprocessing text, it starts from cleaning, filtering, tokenizing, and stemming. The results and analysis stage consists of weighting TF-IDF, processing using both algorithms to find the best cluster validity, visualizing the word and analysing, and doing documentation.

Preprocessing has a very important role in text mining techniques [14]. Preprocessing stages include data preparation, integration, cleaning process, normalization, transformation, complexity reduction, and irrelevant things removal [15]. TF-IDF is a method for evaluating the importance of words in documents. Whether the word is important or not, it depends on the number of times the word appears in a document [16].

2.1. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is an algorithm designed to find clusters and noise in database space [17]. The basic concept of this algorithm is that for each data point in a cluster, an environment with a given radius (Eps) must contain at least a minimum number of points (Minpts) that is the intensity of the environment must exceed several thresholds [18]. The DBSCAN algorithm is started by calculating the distance of the central point (p) to another point using the Euclidean distance and stated in equation 1 [19].

$$D(x_l, p_i) = \sqrt{\sum_{j=i}^{q}(x_{lj} - p_{ij})^2} \qquad (1)$$

2.2. K-Medoids

The K-Medoids algorithm is used to find medoids in a cluster. K-Medoids is stronger than K-Means in finding k as a representative object to minimize the number of data object inequality, reduce noise and outliers [20]. The basic strategy of this algorithm is to find k clusters in n objects first randomly. Each remaining object is grouped with the most similar Medoid. K-Medoids algorithm uses representative objects as representative points in retrieving the average value of objects in each cluster [21]. The distance between objects i and j is calculated using the dissimilarity measurement function, where one of them is the Euclidean Distance Function shown in equation 2 [22]:

$$\mathrm{dii} = \sqrt{\sum_{a=1}^{p}(x_{iu} - x_{iu})^2, i = 1, \ldots\ldots.\ n; j = 1, \ldots.\,n} \qquad (2)$$

From the equation above, Xia is the a-variable of object i (i = 1, ..., n; a = 1, ..., p) and dij is the Euclidean Distance value. The algorithm also calculates the exchange probability of each object with another cluster center using criteria functions such as equation 3:

$$E = \sum_{j=1}^{k}\sum_{\rho\epsilon cj}^{n}|p - 0j| \ldots\ \ldots\ldots\ldots\ldots \qquad (3)$$

The equation 3 above implies that E is the sum of absolute errors for all objects in the dataset; p is the point in the space that represents an object in the Cj cluster, and oj is the object in the Cj cluster.

3. Results And Analysis

At the analysis stage, natural disaster data in Indonesia in 2018 and 2019 were analysed by using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and K-Medoids was used to group data into clusters. The data used in this study were obtained from Twitter data by using crawling techniques. Crawling was performed by using the Python programming language with the keywords natural disasters, earthquakes, earthquakes, tsunamis, floods, landslides, fires, volcanic eruptions, and tornados. The data was taken in the span of 1 January 2018 - 30 September 2019.

Preprocessing Text aims to prepare raw data before the next process is performed. In general, Text Preprocessing is the process of eliminating inappropriate data or changing data to a more suitable form so that it is easy to process. The stages of Text Preprocessing in text mining consist of Tokenizing, Filtering, Stemming, Tagging and Analyzing [23] then the labeling process. The labelled data are calculated to obtain the TF-IDF weights by using the Python programming language. TF-IDF is calculated by eliminating terms with maximum proportion of frequency documents exceeding 90% and minimum frequency document is less than 2% so that it produces each term from the query "natural disasters, earthquakes, earthquakes and tsunamis, floods, landslides, fires, and the volcano erupted".

3.1. Data Grouping Using DBSCAN Algorithm

Twitter text documents that had been weighted using TF-IDF were clustered by using the DBSCAN algorithm with different epsilon (Eps) and minpoints (MinPts) parameters.

3.2. Positive Label Natural Disaster Data

Table 1 is the result of the number of clusters from the clustering experiment with DBSCAN using several different Eps and MinPts parameters.

Table 1. The Cluster Result of Positive Label Natural Disaster Data

No	Eps	MinPts	Silhouette Index	Number of Clusters	Noise	Execution Time (Seconds)
1	0.3	5	0.9371	26	898	20
2	0.3	6	0.9265	22	918	16
3	0.3	7	0.9182	19	936	15
4	0.3	8	0.9182	19	936	12
5	0.3	9	0.9137	18	944	16
6	0.3	10	0.9153	17	953	18
7	0.4	5	0.8547	28	851	12
8	0.4	6	0.8491	26	866	13
9	0.4	7	0.8593	19	905	14
10	0.4	8	0.8593	19	905	12
11	0.4	9	0.8517	18	913	16
12	0.4	10	0.8517	18	913	14
13	0.5	5	0.7613	31	797	13
14	0.5	6	0.7697	26	825	18
15	0.5	7	0.7647	22	850	16
16	0.5	8	0.7760	21	861	13
17	0.5	9	0.7829	19	877	18
18	0.5	10	0.8004	18	886	11
Average execution time (seconds)						14.83

The lower epsilon value, then higher the Silhouette Index value, as well neutral and negative labels. Neutral and negative labels are not displayed in this paper.

3.3. Data Grouping Using the K-Medoids Algorithm

Twitter text documents that had been weighted using TF-IDF were clustered by using the K-Medoids algorithm with different k parameters.

3.4. Positive Label Natural Disaster Data

K-Medoids cluster results for natural disaster data with positive label is shown in Table 2.

Table 2. The Cluster Result of Positive Label Natural Disaster Data

No.	K	Silhouette Index	Number of Clusters	Execution Time (Seconds)
1	2	0.0530	2	169
2	3	0.0380	3	142
3	4	0.0842	4	228
4	5	0.0805	5	220
5	6	0.0687	6	171

No.	K	Silhouette Index	Number of Clusters	Execution Time (Seconds)
6	7	0.0836	7	235
7	8	0.1048	8	207
8	9	0.1208	9	256
9	10	0.0925	10	236
	Average execution time (seconds)			207.11

3.5. Cluster Validity

Cluster validity aims to obtain the best cluster from several experiments that have been carried out using Silhouette Index (SI). The best SI value is the greatest one or close to 1. The comparison of the best SI value from the positive, negative and neutral label natural disaster data from the DBSCAN and K-Medoids algorithm can be seen in Figure 2.

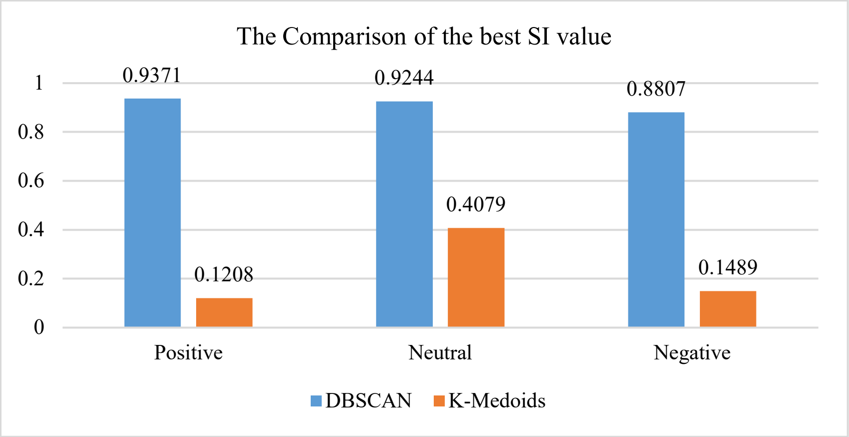

Figure 2. The Comparison of the best SI value

4. Conclusion

The DBSCAN algorithm has the highest Silhouette Index (SI) cluster validity with an average of 0.9140 and average execution time of 83.40 seconds. Meanwhile, the K-Medoids algorithm has a SI value of 0.2258 with an average execution time of 849.93 seconds. The frequency of the word "earthquake" dominated for the positive category, the word "disaster" dominated the negative category, and the word "flood and earthquake" dominated the neutral category. In the positive category table, for further analysis, it contained expressions of empathy or public concern on the earthquake events that happened in Indonesia. On the other hand, in the neutral category table, it contained information about natural disaster events such as floods and earthquakes, and in the negative category table, it contained criticism to the government related to natural disaster management.

References

[1] A. A. Alalwan, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, “Social media in marketing: A review and analysis of the existing literature,” *Telemat. Informatics*, vol. 34, no. 7, pp. 1177–1190, 2017.

[2] F. Anwar, “Perubahan dan Permasalahan Media Sosial,” *J. Muara Ilmu Sos. Humaniora, dan Seni*, vol. 1, no. 1, p. 137, 2017.

[3] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, “Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.

[4] N. Öztürk and S. Ayvaz, “Sentiment analysis on Twitter: A text mining approach to the Syrian

refugee crisis," *Telemat. Informatics*, vol. 35, no. 1, pp. 136–147, 2018.

[5] R. Fattah, "Twitter Text Mining Untuk Informasi Gempa Bumi Menggunakan Tf-Idf Di Indonesia," Universitas Islam Negeri Maulana Malik Ibrahim, 2016.

[6] Fatimahsyam, "Pengintegrasian Pengurangan Risiko Bencana Dengan Pendekatan Mazhab Antropocentris," *Subtantia*, vol. 20, no. March, pp. 49–65, 2018.

[7] Ramadhan and M. Iqbal, "Penerapan Data Mining Untuk Analisis Data Bencana Milik BNPB Menggunakan Algoritma K-Means Dan Linear Regression," *J. Inform. dan Komput.*, vol. 22, no. 1, pp. 567–65, 2017.

[8] N. G. I. Reza, "Penerapan Algoritma Dbscan Untuk Pencarian Trend Topik Pilkada Pekanbaru 2017 Pada Twitter," 2018.

[9] R. N. G. Indah, R. Novita, O. B. Kharisma, R. Vebrianto, S. Sanjaya, T. Andriani, W. P. Sari, Y. Novita, and R. Rahim, "DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru," in *Journal of Physics: Conference Series*, 2019, vol. 1363, no. 1, p. 12001.

[10] S. Dang and P. H. Ahmad, "Text Mining : Techniques and its Application," *Int. J. Eng. Technol. Innov.*, vol. 1, no. 4, pp. 22–25, 2014.

[11] R. Feldman and J. Sanger, *The Text Mining Handbook*. Cambridge University Press, 2006.

[12] Z. Ramadhan, Aditya, Mustakim, Efendi, "Perbandingan K-Means dan Fuzzy C-Means untuk Pengelompokan Data User Knowledge Modeling," *Semin. Nas. Teknol. Informasi, Komun. dan Ind. 9*, pp. 18–19, 2017.

[13] L. Alfi, I. Atastina, and A. Herdiani, "Analisis Dan Implementasi Community Detection Menggunakan Algoritma Dbscan Pada Twitter," *e-Proceeding Eng.*, vol. 5, no. 1, pp. 1469–1476, 2018.

[14] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, "Preprocessing Techniques for Text Mining -An Overview," *Int. J. Comput. Sci. Commun. Networks,* vol. 5, no. 1, pp. 7–16.

[15] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016.

[16] U. Erra, S. Senatore, F. Minnella, and G. Caggianese, "Approximate TF-IDF Based on Topic Extraction from Massive Message Stream Using the GPU," *Inf. Sci. (Ny).*, vol. 292, pp. 143–161, 2015.

[17] X. Ester, Martin, Krigel, Hans-Peter, Sander, Jorg, Xu, "Density-Based Clustering Methods," *Compr. Chemom.*, vol. 2, pp. 635–654, 1996.

[18] M. Patwary, D. Palsetia, A. Agrawal, W. K. Liao, F. Manne, and A. Choudhary, "A new Scalable Parallel DBSCAN Algorithm Using the Disjoint-Set Data Structure," *Int. Conf. High Perform. Comput. Networking, Storage Anal. SC*, 2012.

[19] S. A. D. Budiman, D. Safitri, and I. Dwi, "Perbandingan Metode K-Means dan Metode Dbscan pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang," *J. Gaussian*, vol. 5, no. 4, pp. 757–762, 2016.

[20] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016.

[21] T. Santhanam and T. Velmurugan, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points," *J. Comput. Sci.*, vol. 6, no. 3, pp. 363–368, 2010.

[22] Z. Mustofa and I. S. Suasana, "Algoritma Clustering K-Medoids pada E-Goverment Bidang Information and Communication," *J. Teknol. Inf. dan Komun.*, vol. 9, no. 1, pp. 1–10, 2018.

[23] W. Gata, "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS," *J. Format*, vol. 6, pp. 1–13, 2017.