Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11569
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Performance Comparison of K-Means and DBScan Algorithms for Text Clustering Product Reviews

Fitri Andriyani[1)*], Yan Puspitarani[2)]
[1,2)]Universitas Widyatama, Indonesia
[1)]fitri.andriyani@widyatama.ac.id, [2)]yan.puspitarani@widyatama.ac.id

Submitted : July 25, 2022 | **Accepted** : July 18, 2022 | **Published** : July 25, 2022

Abstract: The purpose of this study was to compare the accuracy performance of the K-Means and DBScan algorithms in clustering product reviews. This comparison evaluated to determine which algorithm is better in terms of accuracy. The two algorithms were chosen because they have different methods of clustering, K-Means uses centroid-based while DBScan uses density-based. Text clustering results can be implemented on e-commerce platforms, marketplaces or product review platforms. This can help customers in deciding what product they will buy. One of the factors that customers have difficulty in determining what product they will buy is the number of reviews that each product has, and the difficulty in concluding the advantages of each product that will be matched their needs or desires. With text clustering, it can be easier and faster for customer to determine whether the product is worth buying or not based on the product reviews they read. The data set used in this study is a review of the Cetaphil Facial Wash product from the Female Daily website. Firstly, data set goes through the Text Pre-Processing stage; then it will be clustered using two algorithms, K-Means and DBScan. After that, the results of the clustering of the two algorithms calculated for their accuracy performance and the performance results obtained. From the results of this study, it concluded that, in the review clustering of Cetaphil Facial Wash products, DBScan has 99.80% accuracy, which higher to compare with K-Means with only has 99.50% accuracy.

Keywords: DBScan, K-Means, Product Review, RapidMiner, Text Clustering,

INTRODUCTION

Today, the purchase of a product or service usually done online. For those who shop online from the market, it is very difficult for them to find the product they need. On the marketplace platform, a search result will show different types of suggested products from relevant brands or stores. This makes it difficult for customers to find products that directly match their desires. Buyers will be attracted to products with high sales and high ratings. Even if a product has many reviews, it is not necessarily helpful because many of the reviews are less relevant and contain spam. Knowing the benefits of a product, shoppers are also used to finding product reviews on specialized review platforms, for example on health and beauty products, namely Female Daily or Sociolla.

On the Female Daily platform, there are different types of skin care and beauty products from different brands and the product has many reviews from other users who have used the product. The platform offers suitable, more consistent and rarely found spam reviews. To improve the customer experience, it is necessary to group the benefits of the products according to the reviews of each product using Text Clustering.

Clustering, which is included in Unsupervised Learning, is one of the analytical methods that works to group objects into the same clusters by judging from the similarity of the data to one another without the information of the class label (Rodriguez et al., 2019). In several studies, algorithms for clustering, especially for product reviews, mostly uses K-Means and DBScan. Both algorithms have high accuracy in carrying out clustering techniques. In research conducted by Laksmanaprabu in 2018 (Lakshmanaprabu et al., 2018), which showed that the accuracy of the K-Means algorithm in the clustering method was 98% in analyzing user reviews, and in research conducted by Mustafa Cataltas in 2020 (Cataltas, Dogramaci, Yumusak, & Oztoprak, 2020) shows that the DBScan algorithm shows good performance in finding product defects from customer reviews. These two algorithms have different characteristics, K-Means can handle large amounts of data and the number

Corresponden author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11569
e-ISSN : 2541-2019
p-ISSN : 2541-044X

of central point needs to initialize from the beginning. While DBScan has the characteristics of being able to accommodate data that has unlimited outliers and in its application there is no need to initialize the number of clusters at the beginning. This study aims to compare the two algorithms and measure accuracy performance in text clustering on product review and to see the grouped words in every cluster using Wordcloud. The application of clustering to both algorithms expected to help the review platform in choosing which algorithm is better in product review clustering.

LITERATURE REVIEW

Several previous studies used the K-Means & DBScan algorithm in their comparison experiments. The results of research in automatic velocity picking analysis show that the accuracy results provided by K-Means are sensitive to cluster values, while the performance of DBScan does not have much influence even though the epsilon value changes (bin Waheed, Al-Zahrani, & Hanafy, 2019). In the results of the research on grouping villages in Central Java using both algorithms, it has a higher silhouette coefficient value on the application of the DBScan algorithm (Dewi et al., 2021). In the application review clustering using the K-Means, Agglomerative, DBScan + Agglomerative algorithm, it shows that the value shown by the DBSCAN + Agglomerative algorithm is the highest (Wei, Lao, Sato, & Han, 2019). In a study to find out product defects in product reviews written by buyers using DBScan (Cataltas et al., 2020) showed good results with the identification of clusters which are groups of product defects. From the results of this study, it can also be found that words often mentioned related to product defects and managed to get these results by using negative reviews from buyers. By using the K-Means and DBScan algorithms on rainfall data, the two algorithms produce different clusters; from the results of the clustering, it founds that K-Means produces more efficient and accurate result (Yan et al., 2020). In a study that compares the performance of the two algorithms for data from customer support, it shows that the results of K-Means have several data points that enter several clusters; the performance of K-Means itself has a high recall value but low precision value. Meanwhile, DBScan has problems in identifying the relevant clusters from the data corpus used (Kästel & Vestergaard, 2019).

From these studies, comparisons between K-Means and DBScan algorithms often made, but there is no set value indicating that one is better than the other. In addition, the use of the dataset also affects the results obtained, by determining the k value for K-Means and determining the Eps value in the DBScan algorithm affects the clustering of the results. In this study, K-Means & DBScan algorithms will compared for the accuracy performance based on the highest accuracy value when it has given different parameter value and will showed the Wordcloud chart to observe commonly mentioned words in each cluster generated by the two algorithms. In this study, we compared the K-Means algorithm with the DBScan algorithm to see the accuracy performance based on the highest accuracy score when various parameter values specified, and displayed a Wordcloud chart to observe the words mentioned in each clusters and show how the words correlate.

METHOD

The method used in this study based on Text Mining, which has several steps to conduct mining form text. Text Mining is a process of finding and extracting valuable knowledge and information from data in the form of text (Jo, 2019), it can be from people's words, text messages, reviews, opinions, transcripts of news shows, transcripts of podcasts, email, newspapers and others (Jayasekara & K.S., 2018). Technological developments also play a role in contributing to the large amount of unstructured data such as digital text data on various social media platforms, messaging apps or other applications that can analyzed and utilized. In text mining, the data used are text and text forms in real life that are unstructured, unshaped and difficult to process through algorithms (Witten, 2004). Text consists of a combination of letters that become a word, form a sentence, become a paragraph and have a certain sentence pattern. To extract information from a text, classification, regression, clustering and association processes carried out as the main processes for data mining (Jo, 2019). The text mining process starts from unstructured data, then undergo long processes, which ultimately information or knowledge obtained from the data that has collected.

The dataset used in this study is a review of Cetaphil Facial Wash on the Female Daily website and contains 3,309 sentences of product reviews extracted using a web scrapper. The data obtained then enters the Text Preprocessing stage, which will generate tokens that will used in the text mining process. In doing text mining, structured data will make it easier and produce the appropriate output. The data source used in this study is unstructured data; therefore, this data needs to process first to produce cleaner, structured and ready data for analysis (Chandrayan & Bamne, 2021). In the preprocessing stage, there are several methods used to generate tokens that will used for the next stage, and it produced 57 list of words that will be using in the text clustering process.

The preprocessing methods used in this research are Tokenize, Case Folding, Stop-Word Removal and Stemming. Tokenization is a process of dividing a sentence or document that is usually separated by spaces or

Corresponden author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11569
e-ISSN : 2541-2019
p-ISSN : 2541-044X

punctuation into a set of words called tokens (Jo, 2019; Shiri, 2004). The generated token then converted at the case folding stage by converting each word in lowercase. This step prevents the occurrence of the same word with different case folding. After that, Stop-Word Removal is performed, this step aims to remove unnecessary words that usually in a form of conjunction or a words that doesn't have a sentiment (Khomsah & Agus Sasmito Aribowo, 2020), such as and, or, but, because and however. These words removed from a text to forming a cleaner and more meaningful data. This process used public stop word dictionary by Gene Diaz in GitHub (Diaz, n.d.) and added some additional words due to words generated by previous step. Next process is Stemming, this step has purpose to extract root word by removing the affix of a word (Magriyanti, 2018) it also has advantage to remove the slang words from the dataset (Khomsah & Agus Sasmito Aribowo, 2020). In this research this process use to stem words like, *membersihkan* become *bersih*; *berbusa* to *busa*; *kerasa* to rasa and much more. Because of the amount data is not large, the stemming dictionary used is only contains 45 words.

The words produced by the previous step will measured in the frequency of occurrence using TF-IDF. The purpose of this process is term weighing to replace the VSM cell value, which was originally the number of occurrences of the term in each document, with the weight value for the term using TF-IDF. Then, to get the TF-IDF value for each term and document, it will calculate the frequency value of the word that occurs in the document, multiplied by the term frequency (TF) by the inverse document frequency (IDF) which represents the number of documents.

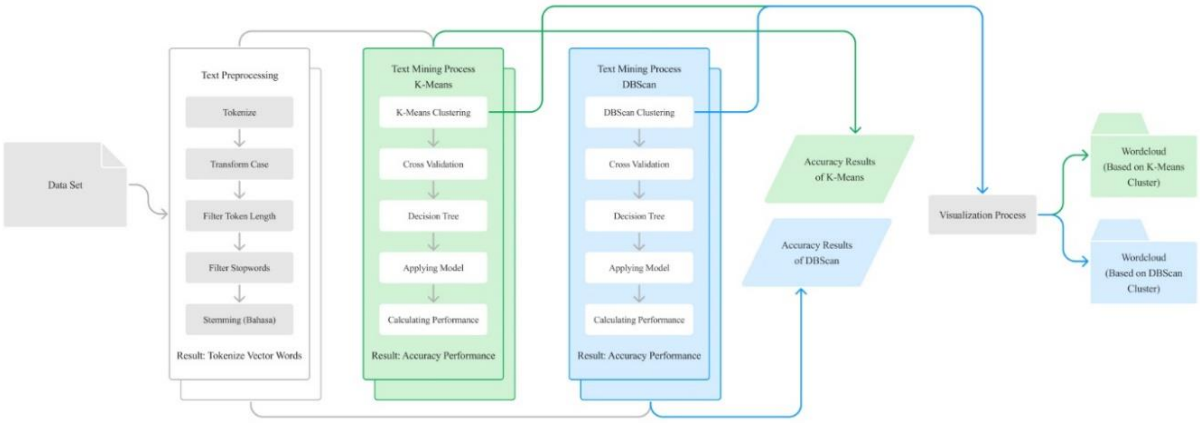

Fig. 1 Text Preprocessing & Text Mining Process

Tokens obtained from the results of Text Preprocessing then go through a text mining process that uses both algorithms, namely K-Means and DBScan. The K-Means algorithm is an algorithm that often used for clustering techniques because of its simplicity and computational efficiency (Kłopotek, 2020). The K-Means algorithm takes data points as input and groups them into k clusters, the result will be a model that will take sample data as input and will group the new data points according to the training model data. K-Means has several components, namely, centroid that is the cluster to formed, k is the number of centroids and data points that represent data that will processed by the K-Means algorithm. Each centroid of the cluster is a set of feature values that define the group it will generate. Each centroid feature will examined for use in interpreting the type of group represented by each cluster qualitatively. To determine the optimal k value, repeated iterations carried out to determine the performance of each k value used. This done to determine which k value has the highest accuracy value.

The next process is to apply the DBScan algorithm in product review clustering. DBScan is a density based clustering algorithm for data that has a more abstract form. The main key in DBScan is that each cluster object must meet the requirements, namely, the distance to neighboring points must match the radius (Eps) that has defined and have a minimum of data points ($MinPts$) within that radius. This means that the cardinality for a range (cluster) must exceed the ($MinPts$) that has been defined (Rehman, Asghar, Fong, & Sarasvady, 2014). There are three types of points, namely core points, borders and outliers that distinguished by the number of points in the radius range (Eps). In determining the optimal value of epsilon and min points, repeated iterations also carried out to determine the performance of which epsilon value has the highest accuracy value. The determination of the epsilon range done by checking the equations of the data used, the range of epsilon value that will be calculated is when the inflection point on the graph occurs, nine range points are taken with a difference value of three. From these results it is determined which epsilon will be used.

From the results of the clustering analysis, the performance calculations performed using cross-validation. Cross-validation is a method of resampling data to evaluate the generalizability of predictive models (Berrar,

Corresponden author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11569
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2019). The cross-validation result obtained in the form of a confusion matrix containing the predictions and real data generated using Decision Tree to find word frequency patterns. These results show accuracy, class recall, and performance accuracy. In addition, to describe the cluster that has formed the cluster will displayed with a Word Cloud graph showing the product review words contained in the cluster. Each word has a different size depending on the size of the word value.

RESULT

In this study, analysis of the resulting cluster formed by giving different values to the parameters of the algorithm performed. The k parameter for the K-Means algorithm and the epsilon parameter for the DBScan algorithm. On Table 1 is the result of calculating the performance of the K-Means algorithm and Decision Tree with a value of k from 2 to 10. The result with the highest accuracy of K-Means with Decision Tree is setting the value of k to 2, giving an accuracy of 99, 50%. In the implementation of the DBScan with Decision Tree algorithm, the average $MinPts$ value is 15, while the Eps value ranges from 0.63 to 0.87. DBScan's highest accuracy result is 0.63 epsilon, forming 4 clusters, and accuracy is 99.80%.

Table 1. Compare k value, accuracy and time for K-Means Algorithm

Value of k	Accuracy	Time (seconds)
2	99.50%	1.177
3	93.22%	1.731
4	91.09%	2.319
5	74.32%	2.735
6	71.80%	3.702
7	64.24%	4.972
8	69.21%	3.360
9	76.05%	4.091
10	67.53%	2.467

Table 2. Confusion Matrix of Highest Accuracy Performance of K-Means Algorithm Using Decision Tree

	True Cluster 1	True Cluster 2	Class Precision
Pred Cluster 1	691	69	90.92%
Pred Cluster 2	532	2290	81.15%
Class Recall	56.50%	97.08%	

Table 1 shows that the highest accuracy value for the data used is if the value of $k = 2$, with an accuracy result of 99.50%. From the table, it can be seen that the greater the value of k, the longer the time required for running the process, but the resulting accuracy tends to be smaller. There was an increase in accuracy again at the value of $k = 8$ and $k = 9$, but at the value of $k = 10$ there was a decrease in accuracy again. The difference from the highest to the lowest accuracy value is 31.97% and the difference in time required is 1,290 seconds.

Table 3. Compare Epsilon value, number of clusters, accuracy and time for DBScan Algorithm

Value of epsilon	Number of clusters	Accuracy	Time (seconds)
0.63	4	99.80%	1.657
0.66	8	99.58%	1.657
0.69	9	98.97%	1.657
0.72	10	97.91%	1.657
0.75	12	95%	1.657
0.78	13	90.73%	1.657
0.81	10	84.39%	1.657
0.84	5	78.11%	1.657
0.87	2	69.18%	1.657

Corresponden author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11569
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4. Confusion Matrix of Highest Accuracy Performance of DBScan Algorithm Using Decision Tree

	True Noise	True Cluster 1	True Cluster 2	True Cluster 3	Class Precision
Pred. Noise	3544	2	1	0	99.92%
Pred. Cluster 1	2	18	0	0	90.00%
Pred. Cluster 2	3	0	6	0	66.67%
Pred. Cluster 3	1	0	0	5	83.33%
Class Recall	99.83%	90.00%	85.71%	100.00%	

Table 2 shows that the highest accuracy value for the data used is if the value of $Eps = 0.63$, with an accuracy result of 99.80% and 4 clusters formed. From the table the value of epsilon is not directly proportional to the number of clusters produced, the value of epsilon = 0.81 has a decrease in the number of clusters that continues to occur until the value of $Eps = 0.87$. However, the greater the epsilon value used, the smaller the accuracy obtained, while for the time spent in executing the algorithm, all epsilon values have the same execution value, which is 1,657 seconds. The difference in accuracy from the highest to the lowest is 30.62%.

From the comparison results above, for the highest accuracy results, the k value for the K-Means taken is 2 with an accuracy of 99.50%, then for DBScan the $Eps = 0.63$ and $MinPts = 15$ values are taken with an accuracy of 99.80%. From these experiments, the accuracy value for DBScan is higher, but the processing time for the algorithm is slower with a difference of 0.48 seconds when compared to the K-Means algorithm. The results of DBScan accuracy are higher when compared to K-Means, in carrying out a clustering analysis on the product review data "Cetaphil Facial Wash". In Fig. 2 and Fig. 3 depicted a Word Cloud graph to find out the words that often mentioned in each cluster of each algorithm.

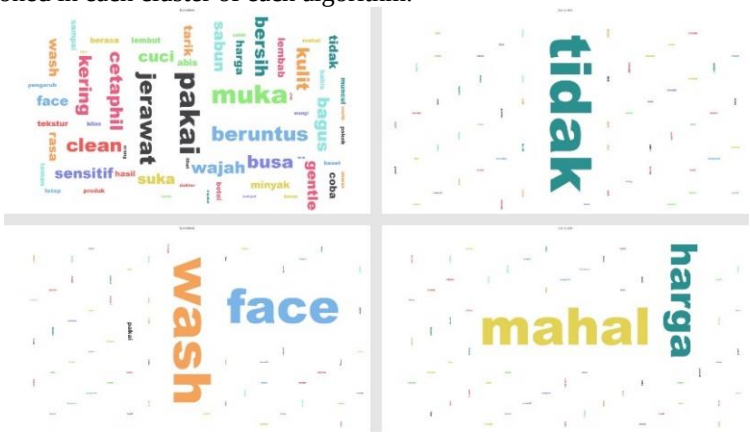

Fig. 2 Wordcloud Results From DBScan Algorithm

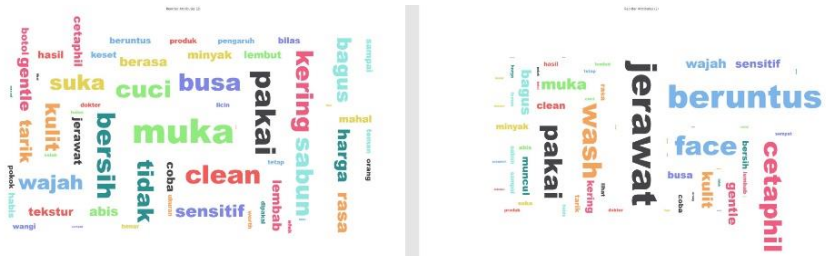

Fig. 3 Wordcloud Results From K-Means Algorithm

DISCUSSIONS

From the results of the study, it founded that the highest accuracy value by calculating the clustering algorithm result, DBScan has higher accuracy score compared to K-Means. DBScan forms four clusters, while K-Means forms two clusters. The formation of the number of clusters is highly dependent on the parameter values given to each algorithm; in K-Means, it means the value of k while DBScan is the value of the epsilon and min scores. Although DBScan has a higher accuracy value and generates more clusters, the DBScan algorithm detects a large number of outliers/noises from the used dataset and in other clusters, i.e. cluster 2 ,3 and 4, Word Cloud Chart results show words that don't match what you want to search first. In comparison, K-Means generates two clusters, but the two clusters have the same weight, the words shown on the Word Cloud graph always have the same value, and there are fewer prominent words than others.

Corresponden author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11569
e-ISSN : 2541-2019
p-ISSN : 2541-044X

CONCLUSION

The application of the K-Means and DBScan algorithms in product reviews for "Cetaphil Product Review" shows that the accuracy of DBScan is higher at 99.80%, while for K-Means it is 99.50%. The difference in parameters is the main reason, with the difference in the parameters of the number of clusters produced being different. To find the most optimal parameter values, it is necessary to carry out continuous iterations, especially for the DBScan parameter with different values given for the epsilon and min points values. Improvement of the results can done with better data processing, because after doing research, the DBScan algorithm detects a lot of noise in the data and the results from other clusters are not satisfactory.

REFERENCES

Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Elsevier. doi:10.1016/B978-0-12-809633-8.20349-X

bin Waheed, U., Al-Zahrani, S., & Hanafy, S. M. (2019). Machine learning algorithms for automatic velocity picking: K-means vs. DBSCAN. *SEG Technical Program Expanded Abstracts*, 5110–5114. doi:10.1190/SEGAM2019-3215809.1

Cataltas, M., Dogramaci, S., Yumusak, S., & Oztoprak, K. (2020). Extraction of Product Defects and Opinions from Customer Reviews by Using Text Clustering and Sentiment Analysis. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, 4529–4534. doi:10.1109/BigData50022.2020.9377851

Chandrayan, S., & Bamne, P. (2021). A brief survey of Text Mining and its applications. *International Journal of Emerging Trends in Engineering Research*, *9*(8), 1190–1195. doi:10.30534/ijeter/2021/26982021

Dewi, C., Siam, E. P., Wijayanti, G. A., Putri, M., Aulia, N., & Nooraeni, R. (2021). Comparison of DBSCAN and K-Means Clustering for Grouping the Village Status in Central Java 2020, *17*.

Diaz, G. (n.d.). Indonesian Stopwords Collection. Retrieved 24 July 2022, from https://github.com/stopwords-iso/stopwords-id/blob/master/stopwords-id.txt

Jayasekara, P. K., & K.S., A. (2018). *Text Mining of Highly Cited Publications in Data Mining*. In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)* (pp. 128–130). IEEE. doi:10.1109/ETTLIS.2018.8485261

Jo, T. (2019). *Text Mining* (Vol. 45). Cham: Springer International Publishing. doi:10.1007/978-3-319-91815-0

Kästel, A. M., & Vestergaard, C. (2019). Comparing performance of K-Means and DBSCAN on customer support queries.

Khomsah, S., & Agus Sasmito Aribowo. (2020). Text-Preprocessing Model Youtube Comments in Indonesian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(4), 648–654. doi:10.29207/resti.v4i4.2035

Kłopotek, M. A. (2020). An Aposteriorical Clusterability Criterion for k-Means++ and Simplicity of Clustering. *SN Computer Science*, *1*(2), 1–38. doi:10.1007/S42979-020-0079-8/TABLES/15

Lakshmanaprabu, S. K., Shankar, K., Gupta, D., Khanna, A., Rodrigues, J. J. P. C., Pinheiro, P. R., & de Albuquerque, V. H. C. (2018). Ranking analysis for online customer reviews of products using opinion mining with clustering. *Complexity*, *2018*. doi:10.1155/2018/3569351

Magriyanti, A. A. (2018). ANALISIS PENGEMBANGAN ALGORITMA PORTER STEMMING DALAM BAHASA INDONESIA.

Rehman, S. U., Asghar, S., Fong, S., & Sarasvady, S. (2014). *DBSCAN: Past, present and future*. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)* (pp. 232–238). IEEE. doi:10.1109/ICADIWT.2014.6814687

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. doi:10.1371/journal.pone.0210236

Shiri, A. (2004). Introduction to Modern Information Retrieval (2nd edition). *Library Review*, *53*(9), 462–463. doi:10.1108/00242530410565256

Wei, Y., Lao, Y., Sato, Y., & Han, D. (2019). Product-review classification combining multiple clustering algorithms. *ACM International Conference Proceeding Series*, 133–136. doi:10.1145/3338188.3338211

Witten, I. H. (2004). Text mining. *The Practical Handbook of Internet Computing*, 14-1-14–22. doi:10.1201/9780203507223

Yan, N., Wu, B., Chang, S., -, al, Pamuji, G. C., & Rongtao, H. (2020). A Comparison study of DBScan and K-Means Clustering in Jakarta rainfall based on the Tropical Rainfall Measuring Mission (TRMM) 1998-2007. *IOP Conference Series: Materials Science and Engineering*, *879*(1), 012057. doi:10.1088/1757-899X/879/1/012057

Corresponden author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.