Hierarchy Clustering Implementation on YouTube's Top Data

Sekar Mulyani(✉), Iin Fatonah, M. Wildan Santosa, Imam Tahyudin, Andi Dwi Riyanto, and Dhanar Intan Surya Saputra

Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia
sekar.mulyani.si.c@gmail.com, iinf28204@gmail.com, himeka.rin09@gmail.com, {imam.tahyudin,andi, dhanarsaputra}@amikompurwokerto.ac.id

Abstract. Clustering is a method or process of grouping datasets into various clusters to produce variations in smaller clusters. Clustering has broad application fields such as data concept construction, pattern recognition, web search, simplification, security, and several other areas. Clustering methods are classified into two types, hierarchies and partitions. The hierarchical clustering method defines the cluster hierarchy by separating and combining them, whereas the partitioning method involves defining and evaluating sections based on criteria. Thus, the selected clustering algorithm must be efficient. This article focuses on clustering algorithms for obtaining and processing YouTube Channel Top Data.

Keywords: Algorithm · Clustering · Hierarchy Clustering · YouTube

1 Introduction

The hierarchical clustering algorithm can be described as a tree in which the data is. The root is a single cluster with all the data, whereas the leaves are individual data objects. There is an intermediary group of subsets of the data between the source (the root) and the leaf [1, 2]. The main idea of a clustering hierarchy is to make "clusters of clusters" go up to build a tree [3]. In order to create hierarchical groups, there are primarily two conceptual techniques. Divisional clustering begins with all the data in a single, huge group and then breaks it down until each piece of data is in its own small group [4].

Data mining enables the extraction of knowledge from historical data and forecasting of future events. Data mining includes clustering, the process of grouping things into groups whose members are related in some way [5]. Clustering can also be defined as grouping data into classes or clusters so that objects in collections have on [7, 8]. Data is a collection of facts, images, or behavior that will be studied manually so that it can predict things that will happen in the future [6]. Data Mining is very important and needs to be done especially for processing extensive data, facilitating the activity of recording a transaction, pattern, or behavior, and for data warehousing processes to provide accurate information for users.

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 437–445, 2022.
https://doi.org/10.2991/978-94-6463-084-8_36

Like Youtube, which applies Data Mining to processing data obtained from its users, Youtube is a web-based application that allows users, viewers, or creators to upload, watch, and share videos [9]. Youtube implements Data Mining; because of the large number of Youtube users, the number of videos that continue to be uploaded, and also the user sees specific videos that they want to watch, automatically have been recorded by the database owned by Youtube. This data collection is collected and processed, which in the end, into Big Data. Big Data is a term that describes a large volume of data; generally, the data is on the internet or can be accessed online; both structured and unstructured [10]. From the data that has been collected and processed, Youtube then presents recommended videos to users according to videos that are often watched or even liked. Using Hierarchy Clustering, YouTube Channel Top Data will be obtained.

The hierarchical clustering method can be done through single Linkage hierarchical clustering, complete Linkage method, and Average Linkage method [11]. Hierarchical grouping can also be done through Cure (Clustering Using REpresentatives), BIRCH Balanced Iterative Reducing and Clustering using Hierarchies, ROCK (RObust Clustering using Links), CHAMELEON Algorithm, Linkage Algorithms, Leaders–Subleaders, bisecting k-Means, all types of hierarchical grouping will remain important for a long time [1, 12].

Clustering of data the challenge of grouping N Data points into K Groups to reduce intra-group "difference" metrics, like the Sum of the squares of cluster centers, is known as the hierarchical clustering approach. Nested cluster sequences, with single-point clusters at the bottom and all-inclusive clusters at the top, can likewise be produced through hierarchical clustering.

2 Methods

In this article, we will explain some of the definitions and definitions of clustering and clustering hierarchies and the methods in the clustering hierarchy.

2.1 Clustering

Clustering converts a group of abstract objects into a class of similar things. A group of data objects can be treated as a group. It is a defined technique for grouping a set of objects, called clusters, based on their characteristics, thereby finding structures in unlabeled data.

Most people can also say that grouping is now a part of every area of life. To illustrate the group, let's consider an example of a hospital management system in which each patient is a data object. Patients with similar symptoms may be placed in one general group because they may need available treatment. This makes it easier for doctors to treat large numbers of patients in a short time.

Different groups can be compared to clusters with patients in one group as data items for one collection. The basis of clustering is to partition large data sets into various similarity-limited domains to make study and analysis; things are simpler because clusters are formed in such a way that the similarity between clusters (similarity between clusters) is minimum and intra-cluster similarity (similarity within clusters) maximum.

There are several methods for the clustering process. However, most methods are part of two broadly defined categories for clustering methodologies, hard clustering, and soft clustering. Hard clustering is a technique where each data element of a data set can be part of only one cluster. Also called an exclusive cluster because there will be no overlapping clusters. One of the most common complex clustering methods is the K-Means method.

2.2 Hierarchical Grouping Algorithm

The hierarchical clustering algorithm generates hierarchical clusters, and the classification of sets depends on whether the hierarchical decomposition is formed in a bottom-up or top-down style. This results in nested clusters starting with one object as a cluster to a single cluster with all objects or vice versa.

In general, merging and splitting are carried out by inferring minimum costs. In the existing literature, the clustering algorithm hierarchy is often represented by a dendrogram. This simple pictorial representation shows the different stages/sequences in which clusters are divided or combined. One of the most attractive features of hierarchical clustering is that the number of clusters to be formed is not fixed at the initial level. Therefore, the desired number of clusters can be created in this clustering, making it a flexible approach.

However, the main drawback is that once the merging or splitting process occurs, making adjustments within a cluster is difficult. So, if the method is not selected in a good way, it can lead to low-quality clusters. Another problem with this type of clustering is that there is no globally defined function for creating clusters. Since cluster formulation is a local step described at each stage, no defined process can lead to proper cluster formation. Hierarchical clustering supports storing data points in the form of a proximity matrix. This article aims to discuss the hierarchical clustering algorithm in its implementation using K-NIME and several methods in the clustering hierarchy.

The following are the steps taken to implement the clustering hierarchy:

1. Prepare data where the data used is numeric type data so that it can be used for distance calculations.
2. Calculate the distance between paired data in the dataset. The calculation method can be selected based on the data.
3. Create a dendrogram from the distance matrix using a specific linkage method. We can also try several linkage methods and then choose the best dendrogram.
4. Determine where to cut the tree (with a specific value). This is the stage where the cluster will be formed.
5. Interpreting the dendrogram that has been obtained.

CURE (Clustering Using Representatives). CURE is the agglomerative hierarchical clustering algorithm CURE (Clustering Using Representatives) balances the all-point and centroid approximations. Dataset partitioning is a hierarchical clustering approach called CURE. To manage huge databases, partitioning and random sampling are combined here [13].

This procedure involves partitioning a random sample from the dataset, followed by partial clustering of each partition. In a second pass, the partial clusters are once more grouped to create the desired clusters. The experiment shows that CURE produces clusters of substantially higher quality than those produced by other algorithms.

BIRCH (Balanced Iterative Reduction and Grouping using Hierarchy). Agglomerative hierarchical clustering technique BIRCH (Balanced Iterative Reduction and Grouping using Hierarchy) is ideally suited for huge databases. The number of I/O operations has been kept to a minimum with this technique. The BIRCH method begins by using a tree structure to partition objects hierarchically, then employing a different clustering algorithm to enhance clusters [14].

This method attempts to achieve the highest quality clustering with the available resources, such as memory and time limits, by progressively and dynamically grouping the incoming data points. To produce clusters of the highest quality, the BIRCH process primarily needs to complete four phases. The terms feature clustering and feature tree clustering (CF Trees), which are used to condense the cluster representation, are introduced in this procedure. For hierarchical clustering, clustering features are stored in CF Trees, which are tall, balanced trees.

ROCK (Robust Clustering Using Links). A potent agglomerative hierarchical clustering technique based on the concept of links is called ROCK (Robust Clustering Using Links). Large data sets can be handled with it as well. ROCK employs linkages between data points rather than their separation while combining data points [12].

This approach works best for categorical and boolean data properties. The number of points from various clusters that have the same neighbors determines how comparable a cluster is according to this technique. Compared to conventional methods, ROCK generates higher-quality clusters and has strong scalability characteristics.

Link Algorithm: The linking algorithm is a hierarchical agglomeration method that considers cluster merging based on the distance between clusters. The Single-Link (S-Link), Average-Link (Ave-Link), and Complete-Link Algorithms are the three most important Link Kinds (Comlink). In Single-Link, the separation between any two subgroups is equal to their shortest separation distance [12].

The distance between two subgroups in an Average Link (Ave-Link) is the average distance between them, whereas in a Complete Link, it is the distance that matters the most. Single Links (S-Link) have trouble coping with extreme differences in cluster density and are vulnerable to outliers. On the other hand, it exhibits complete insensitivity to the size and shape of the cluster.

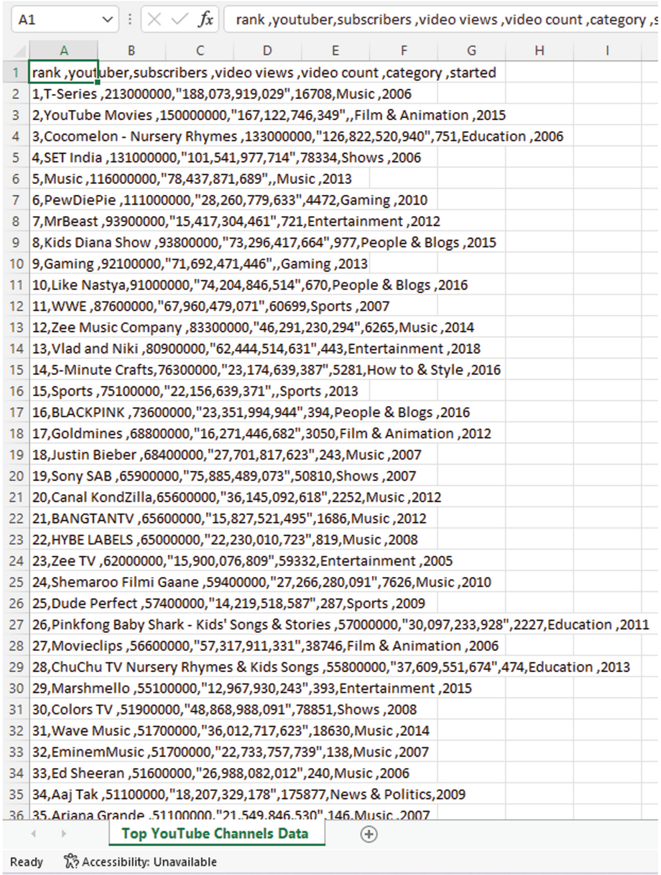
The mean relationship is sensitive to the shape and size of the cluster. Thus, it can quickly fail when the cluster has a complex shape that departs from a hyperspherical form. Full links are not significantly affected by outliers but can break up large groups and have problems with convex shapes.

Bisecting K-Means (BKMS). In the context of document grouping, Bisecting K-Means (BKMS) is a divisional hierarchical clustering algorithm [6]. K-means is a division algorithm that consistently selects the partition with the greatest overall similarity, determined by the pairwise similarity of all the points in a cluster [13].

Halving the K-Means strategy frequently outperforms the traditional K-Means and Agglomerative clustering procedures, and the procedure usually does not end until the necessary number of clusters is obtained as described. The BKMS is low computational cost is a benefit. For grouping huge texts, BKMS was found to perform better than the K- Means (KMS) agglomerative hierarchical method.

3 Results and Discussion

This dataset contains the Top Youtube channels worldwide, which we will analyze using K-NIME with the Clustering hierarchy method; this method is very suitable for clustering. We will enter the dataset image below into the K-NIME application in CSV format (Comma Separated Values) via a reader file that can be opened using the Microsoft Excel application, showing in Fig. 1. CSV is a data format in the database where each record is separated by a comma (,) or a semicolon. In addition to being straightforward,



```
rank ,youtuber,subscribers ,video views ,video count ,category ,s
```

```
rank ,youtuber,subscribers ,video views ,video count ,category ,started
1,T-Series ,213000000,"188,073,919,029",16708,Music ,2006
2,YouTube Movies ,150000000,"167,122,746,349",,Film & Animation ,2015
3,Cocomelon - Nursery Rhymes ,133000000,"126,822,520,940",751,Education ,2006
4,SET India ,131000000,"101,541,977,714",78334,Shows ,2006
5,Music ,116000000,"78,437,871,689",,Music ,2013
6,PewDiePie ,111000000,"28,260,779,633",4472,Gaming ,2010
7,MrBeast ,93900000,"15,417,304,461",721,Entertainment ,2012
8,Kids Diana Show ,93800000,"73,296,417,664",977,People & Blogs ,2015
9,Gaming ,92100000,"71,692,471,446",,Gaming ,2013
10,Like Nastya,91000000,"74,204,846,514",670,People & Blogs ,2016
11,WWE ,87600000,"67,960,479,071",60699,Sports ,2007
12,Zee Music Company ,83300000,"46,291,230,294",6265,Music ,2014
13,Vlad and Niki ,80900000,"62,444,514,631",443,Entertainment ,2018
14,5-Minute Crafts,76300000,"23,174,639,387",5281,How to & Style ,2016
15,Sports ,75100000,"22,156,639,371",,Sports ,2013
16,BLACKPINK ,73600000,"23,351,994,944",394,People & Blogs ,2016
17,Goldmines ,68800000,"16,271,446,682",3050,Film & Animation ,2012
18,Justin Bieber ,68400000,"27,701,817,623",243,Music ,2007
19,Sony SAB ,65900000,"75,885,489,073",50810,Shows ,2007
20,Canal KondZilla,65600000,"36,145,092,618",2252,Music ,2012
21,BANGTANTV ,65600000,"15,827,521,495",1686,Music ,2012
22,HYBE LABELS ,65000000,"22,230,010,723",819,Music ,2008
23,Zee TV ,62000000,"15,900,076,809",59332,Entertainment ,2005
24,Shemaroo Filmi Gaane ,59400000,"27,266,280,091",7626,Music ,2010
25,Dude Perfect ,57400000,"14,219,518,587",287,Sports ,2009
26,Pinkfong Baby Shark - Kids' Songs & Stories ,57000000,"30,097,233,928",2227,Education ,2011
27,Movieclips ,56600000,"57,317,911,331",38746,Film & Animation ,2006
28,ChuChu TV Nursery Rhymes & Kids Songs ,55800000,"37,609,551,674",474,Education ,2013
29,Marshmello ,55100000,"12,967,930,243",393,Entertainment ,2015
30,Colors TV ,51900000,"48,868,988,091",78851,Shows ,2008
31,Wave Music ,51700000,"36,012,717,623",18630,Music ,2014
32,EminemMusic ,51700000,"22,733,757,739",138,Music ,2007
33,Ed Sheeran ,51600000,"26,988,082,012",240,Music ,2006
34,Aaj Tak ,51100000,"18,207,329,178",175877,News & Politics,2009
35,Ariana Grande ,51100000,"21,549,846,530",146,Music ,2007
```

Fig. 1. YouTube Dataset

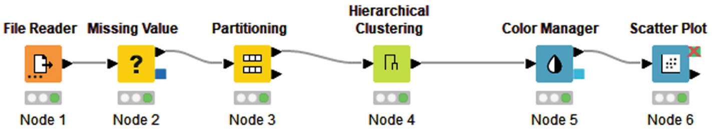
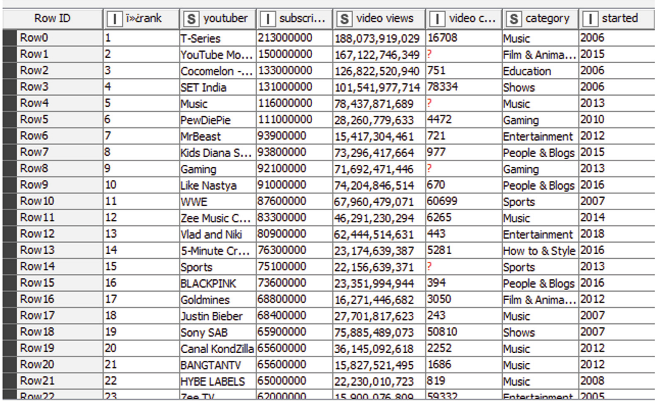


Fig. 2. Hierarchical Clustering Workflow Model



Row ID	I ï»¿rank	S youtuber	I subscri...	S video views	I video c...	S category	I started
Row0	1	T-Series	213000000	188,073,919,029	16708	Music	2006
Row1	2	YouTube Mo...	150000000	167,122,746,349	?	Film & Anima...	2015
Row2	3	Cocomelon -...	133000000	126,822,520,940	751	Education	2006
Row3	4	SET India	131000000	101,541,977,714	78334	Shows	2006
Row4	5	Music	116000000	78,437,871,689	?	Music	2013
Row5	6	PewDiePie	111000000	28,260,779,633	4472	Gaming	2010
Row6	7	MrBeast	93900000	15,417,304,461	721	Entertainment	2012
Row7	8	Kids Diana S...	93800000	73,296,417,664	977	People & Blogs	2015
Row8	9	Gaming	92100000	71,692,471,446	?	Gaming	2013
Row9	10	Like Nastya	91000000	74,204,846,514	670	People & Blogs	2016
Row10	11	WWE	87600000	67,960,479,071	60699	Sports	2007
Row11	12	Zee Music C...	83300000	46,291,230,294	6265	Music	2014
Row12	13	Vlad and Niki	80900000	62,444,514,631	443	Entertainment	2018
Row13	14	5-Minute Cr...	76300000	23,174,639,387	5281	How to & Style	2016
Row14	15	Sports	75100000	22,156,639,371	?	Sports	2013
Row15	16	BLACKPINK	73600000	23,351,994,944	394	People & Blogs	2016
Row16	17	Goldmines	68800000	16,271,446,682	3050	Film & Anima...	2012
Row17	18	Justin Bieber	68400000	27,701,817,623	243	Music	2007
Row18	19	Sony SAB	65900000	75,885,489,073	50810	Shows	2007
Row19	20	Canal KondZilla	65600000	36,145,092,618	2252	Music	2012
Row20	21	BANGTANTV	65600000	15,827,521,495	1686	Music	2012
Row21	22	HYBE LABELS	65000000	22,230,010,723	819	Music	2008
Row22	23	Zee TV	62000000	15,900,076,809	59332	Entertainment	2005

Fig. 3. The incoming data contains seven variables

this format can be opened in several text editors, including Notepad, Wordpad, and Microsoft Excel.

From the Workflow that we built to get the results, Fig. 2. We can conclude that this Hierarchical Clustering method starts with all data points in one large cluster, and the most different data points are divided into sub-clusters until each cluster consists of precisely one data point. Here is our solution, using the K-NIME application.

From the results of the model we created, in the File Reader, there are several columns such as ranking, YouTuber name or YouTube Channel, subscribers, video views, and others (Fig. 3).

The Clustering Hierarchy from the following picture results from grouping based on the highest subscribers.

Figure 4 Dendrogram or Grouping where the X-axis is the videos view and the Y-axis is the subscribers.

In Fig. 5, Distance or Grouping, the X-axis is subscribers, and the Y-axis is video views.

Figure 6 shows the cluster as the X-axis in the video view: Blue and Red are subscribers, and Yellow is the class cluster.

4 Conclusion

Clustering is the process or method of grouping datasets into various clusters to make the variation within the clusters tiny. In contrast, the hierarchical clustering algorithm can be described as a tree in which the data is. A clustering algorithm also has a critical role in

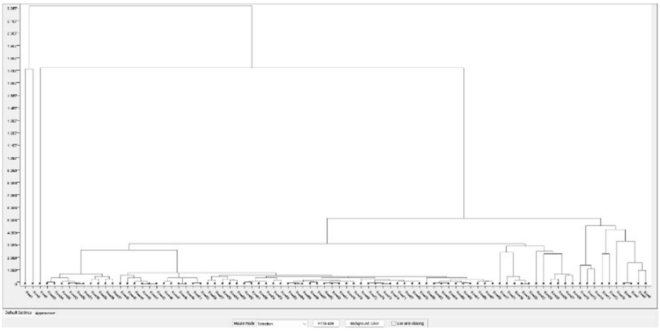


Fig. 4. Dendrogram Display

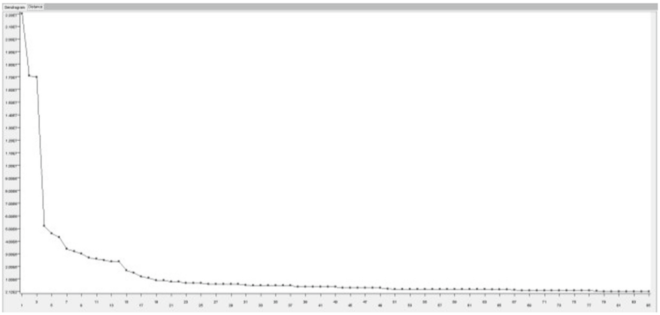


Fig. 5. Display Distance

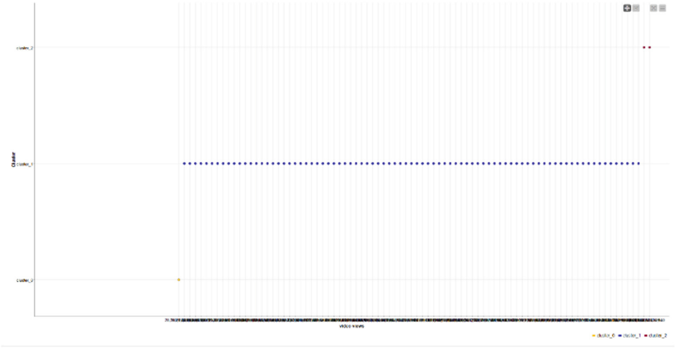


Fig. 6. Cluster View

data mining. Several hierarchical clustering algorithms generate hierarchical clusters and cluster classification. The formation of hierarchical decomposition might be top-down or bottom-up.

Several methods in the first clustering hierarchy are CURE, BIRCH, ROCK, and Bisecting K-means (BKMS). For the dataset that we use regarding the most significant number of subscribers and viewers, we analyze and implement it using K-NIME with the clustering hierarchy method. This method is very suitable for grouping clusters with a more significant number.

Some of the advantages of the clustering hierarchy algorithm:

- Able to describe the closeness between data with the dendrogram
- Quite easy to make
- Can determine the number of clusters formed after the dendrogram is formed.

Some disadvantages of the hierarchical clustering algorithm:

- Cannot analyze categorical data directly
- Not intended to produce an absolute optimal number of clusters
- Sensitive to data that has a different scale, so the data needs to be normalized/standardized first.

References

1. T. Zhang., R. Ramakrishnan., M. Livny.: BIRCH: A New Data Clustering Algorithm and Its Applications. Data Min. Knowl. Discov., vol. 182, no. 1, pp. 141–182 (1997).
2. Y. Rani.,H. Rohil.: A study of hierarchical clustering algorithms. Int. J. Inf. Comput. Technol., vol. 3, no. 11, pp. 1225–1232 (2013).
3. K. A. Wijaya., D. Swanjaya.: Integrasi Metode Agglomerative Hierarchical Clustering dan Backpropagation Pada Model Peramalan Penjualan, In: Seminar Nasional Inovasi Teknologi, pp. 132–141 (2021). Available: https:// proceeding.unpkediri.ac.id/index.php/inotek/article/view/109 2/703
4. Vijaya., S. Aayushi., R. Bateja.: A Review on Hierarchical Clustering Algorithms. Journal of Engineering and Applied Sciences, vol. 12, no. 24. pp. 7501–7507 (2017).
5. S. B. Musa., A. B. Kaswar., Supria., S. Sari.: Document Clustering by Dynamic Hierarchical Algorithm Based on Fuzzy Set Type-II From Frequent Itemset. J. Ilmu Komput. dan Inf. (Journal Comput. Sci. Information), vol. 9, no. 2, p. 2016, (2016). [Online]. Available: http:// dx.doi.org/https://doi.org/10.21609/jiki.v9i2.383
6. P. Shetty., S. Singh.: Hierarchical Clustering: A Survey," Int. J. Appl. Res., vol. 7, no. 4, pp. 178–181 (2021). doi: https://doi.org/10.22271/allresearch.2021.v7.i4c.8484.
7. F. Kurniawan Medium.com Pemanfaatan Big Data oleh Youtube yang Menerapkan Data Mining [Utilization of Big Data by Youtube That ApplyDataMining]. https://www.medium.com/@fandikurniawan988/pemanfaatan-big-data-oleh-youtube-yang-menerapkan-data-mining-4b145c1519a5, last accessed 2022/07/12.
8. S. W. Handani., D. I. S. Saputra., R. M. Arino., G. F. A. Ramadhan.: Sentiment Analysis for Go-Jek on Google Play Store. J. Phys. Conf. Ser., vol. 1196, no. 1 (2019).
9. D. I. S. Saputra., I. Setyawan.: Virtual YouTuber (VTuber) Sebagai Konten Media Pembelajaran Online [Virtual YouTuber (VTuber) As Online Learning Media Content. In: Seminar Nasional Sistem Informasi dan Teknologi (SISFOTEK) ke 5 Tahun 2021, pp. 14–20 (2021).
10. D. I. S. Saputra., I. Tahyudin., D. Mustofa., Hartanto., T. Mahardianto.: Open Big Data for Indonesian Biodiversity Based on an Online Crowdsourcing Platform. Seybold Rep. J., vol. 17, no. 5, pp. 161–171 (2022).

11. E. Hartini.: Metode Clustering Hirarki [Hierarchical Clustering Method. Pus. Pengemb. Teknol. Inf. dan Komputasi BATAN, vol. 1, pp. 1–11 (2014).
12. F. Murtagh., P. Contreras.: Methods of Hierarchical Clustering. Int. Encycl. Stat. Sci., no. April, (2011). doi: https://doi.org/10.1007/978-3-642-04898- 2.
13. D. T. Utari., D. S. Hanun.: Hierarchical Clustering Approach for Region Analysis of Contraceptive Users. EKSAKTA J. Sci. Data Anal., vol. 2, no. 2, pp. 99–108 (2021). doi: https://doi.org/10.20885/eksakta.vol2.iss2.art3.
14. L. Sousa.: Variability analysis of the hierarchical clustering algoritms and its implication on consensus clustering. Int. J. Adv. Eng. Res. Sci., vol. 4, no. 6, pp. 118–131 (2017). doi: https://doi.org/10.22161/ijaers.4.6.14.