


Application of K-Means Algorithm for Segmentation Analysis of Youtube Viewers in Indonesia

Ryan Artanto Halim¹, Heny Pratiwi^{2*}, Azahari³

^{1,2*,3}STMIK Widya Cipta Dharma, Samarinda, Indonesia

| Article Info | ABSTRACT |
|--|--|
| Keywords: K-Means, Audience segmentation, YouTube, Clustering | The application of K-Means as a clustering method in segmentation analysis is common. However, academic research on YouTube audience segmentation in Indonesia is still limited. YouTube audiences in Indonesia are diverse, ranging from entertainment, education, to news, so more in-depth analysis is needed to identify user segments more specifically. YouTube audience segmentation can provide a deeper understanding of people's video consumption behavior. This understanding can help content creators and digital industry players develop more effective content strategies. K-Means was chosen as the clustering method in this study because it can group YouTube viewers in Indonesia based on their interaction patterns with YouTube content. In addition, K-Means' ability to handle large data is suitable for segmenting platforms with a large number of users such as YouTube. This research uses three main features, namely views, duration, and engagement rate to group viewers into five clusters. Cluster evaluation using Silhouette Score (0.3445), Davies-Bouldin Index (0.9576), and Calinski-Harabasz Index (481.4730) shows that the resulting segmentation is of good quality. The analysis shows that there are differences in video consumption patterns across clusters, reflecting variations in viewer preferences and engagement levels. |
| This is an open access article under the CC BY-NC license  | Corresponding Author: Heny Pratiwi STMIK Widya Cipta Dharma, Samarinda, Indonesia henypratiwi@wicida.ac.id |

INTRODUCTION

The application of K-Means as an unlabeled data clustering method has been done in several previous studies involving data mining [1] and machine learning. K-Means is widely used due to several factors such as its simplicity, efficiency in handling large data, computational speed, ease of interpretation of results and its ability to automatically group data based on their similarities [2][3] making it suitable for segmentation analysis on social media platforms such as YouTube where the number of users can reach more than 2 billion globally [4].

YouTube has the advantage of content diversity and video duration, allowing viewers to further explore topics of interest and utilize their time more effectively than other video platforms [5]. The high popularity and wide variety of content offered makes segmentation analysis on this platform a great way to understand video consumption patterns and user behavior more specifically.

As a country that ranks fourth as the country with the highest number of active YouTube users [6], Indonesia has great potential as a target market for content creators and digital industry players. However, academic research that addresses the segmentation of YouTube audiences in Indonesia, particularly in general, is still lacking. YouTube viewers in Indonesia have very varied content preferences, ranging from entertainment, education, to news, making segmentation more complex but also more interesting to study [7]. In addition, the understanding gained from segmenting the Indonesian YouTube audience can also help content creators and digital industry players in developing effective content strategies to reach the potential of the digital market in Indonesia.

Several previous studies have addressed YouTube-based segmentation analysis with various approaches. Sitanggang, et al [8] developed a YouTube video classification system using a combination of K-Means and Support Vector Machine (SVM) algorithms, where K-Means is used to cluster video metadata before being classified with SVM. Widjaja and Oetama [9] applied K-Means to cluster trending videos on YouTube United States based on the number of views, likes, and dislikes, using Elbow Method to determine the best number of clusters. Meanwhile, Perdana, et al [10] applied K-Means in customer segmentation in Alfragift e-commerce based on user characteristics such as age, gender, and purchase patterns.

Different from previous studies focusing on video metadata or e-commerce customer segmentation, this research applies K-Means to group YouTube viewers based on views, duration, and audience interaction with videos. In addition, this research focuses on the Indonesian region of YouTube instead of the United States. This research focuses on understanding the YouTube audience in Indonesia with the aim of helping content creators develop more effective content strategies.

METHODS

This research was conducted with a data mining approach. Data mining is the process of exploring and analyzing large amounts of data to find patterns, relationships, or valuable information that was previously unknown [11]. In this research, data mining is applied through several main stages, namely data collection, data pre-processing, clustering, and result evaluation.

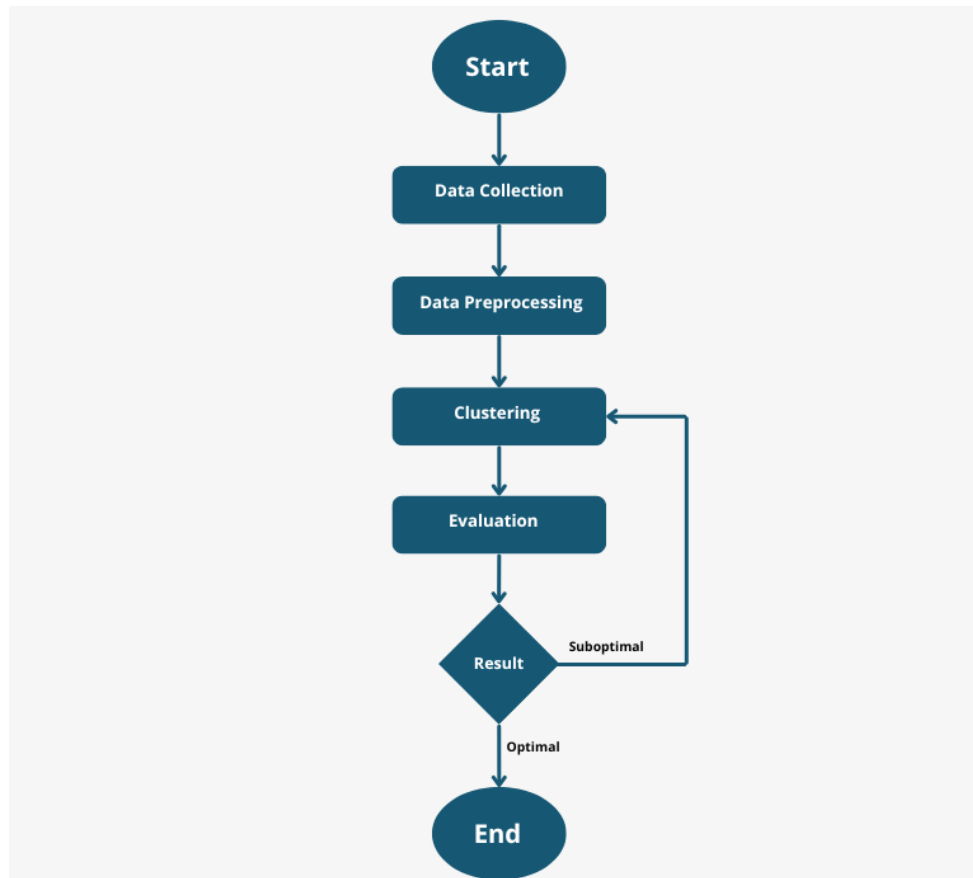


Figure 1: flow

Data Collection

The data in this study was obtained through the YouTube API. The data collected includes:

- Video trending on March 26-28, 2025.
- Videos are relevant to the trending video category to get a wider variety.
- Popular videos from January 1 - March 27, 2025 to see more stable interaction patterns.

Each video collected has features such as number of views, number of likes, number of comments, video duration, engagement rate. Engagement rate is calculated using the formula :

$$\text{Engagement Rate} = \left(\frac{\text{Like} + \text{Comment}}{\text{View}} \right) \times 100\% \quad (1)$$

Description:

Likes = number of likes on the video

Comment = number of comments on the video

View = number of video views

Data Pre-processing

Data pre-processing was conducted to ensure data quality before further analysis. The steps applied include:

Data cleaning

Removing duplicate data and handling missing or invalid values. This resulted in 941 unique samples that reflect the trends of Indonesian audiences.

Outlier Handling and Data Transformation

Log transformation is applied to views, likes, comments, and duration to reduce skewness and suppress the influence of outliers. Then additional outlier handling was performed on views and likes using winzoring due to the high number of outliers even after log transformation. After that, an additional transformation is performed on likes and comments using Yeo-Johnson transformation to overcome the skewness that is still high even after log transformation.

Winzoring the engagement rate before transformation to reduce the impact of extreme outliers. After handling the outliers, the Yeo-Johnson transformation is applied to the engagement rate.

Data Normalization

StandardScaler is used to ensure features have a mean ≈ 0 and standard deviation ≈ 1 for optimal clustering algorithm performance.

Clustering YouTube Viewers

The clustering process is done using the K-Means algorithm, which is one of the unsupervised learning methods to group data based on similarity of features. The clustering steps include:

1. Determine the optimal number of clusters using the Elbow Method and Silhouette Score.
2. Run K-Means with a predetermined number of clusters.
3. Analyze the characteristics of each cluster based on the average value of key features such as views, duration, and engagement rate.

In this study, the clustering process will be repeated using a different combination of features if the evaluation results show that the resulting clusterization is not good.

Evaluation of Clustering Results

To ensure the quality of the clusterization, an evaluation was conducted using three key metrics:

1. Silhouette Score: Measures how well the data is grouped in their respective clusters.
2. Davies-Bouldin Index: Measures the degree of separation between clusters (the smaller the better).
3. Calinski-Harabasz Index: Measures the density and separation between clusters (the bigger the better).

If the evaluation results show that the resulting clusterization is not good enough, then the clustering process will be repeated using a different combination of features. The evaluation results show that the combination of view, duration, and engagement rate features with the optimal number of clusters $k=5$ produces the most optimal clustering compared to other combinations.

RESULTS AND DISCUSSION

The following are the results of clusterization evaluation based on view, duration, and engagement rate features.

- Silhouette Score: 0.3445 (Semakin tinggi semakin baik)
- Davies-Bouldin Index: 0.9576 (Semakin rendah semakin baik)
- Calinski-Harabasz Index: 481.4730 (Semakin tinggi semakin baik)

Evaluation Results

The Silhouette Score of 0.3445 indicates that the cluster structure is quite good. Most of the data falls within the appropriate clusters, although it is possible that some data points are near the boundaries between clusters. The Davies-Bouldin Index of 0.9576 indicates that the clusters are fairly well separated, but there is still some similarity between the groups.

The Calinski-Harabasz Index of 481.4730 shows that clustering is quite effective in separating groups of viewers based on their video consumption and engagement patterns. Overall, the evaluation results show that clustering using K-Means based on view, duration, and engagement rate features with the optimal number of clusters $K=5$ has good performance and can be used as a basis for analyzing YouTube audience segmentation in Indonesia.

Data Distribution in Each Cluster

The distribution of the number of views, video duration, and engagement level in each cluster is presented in the following table:

Table 1. Data Distribution in Each Cluster

| Cluster | Number of Data Points | Views (Mean \pm Std) | Duration (Mean \pm Std) | Engagement Rate (Mean \pm Std) |
|---------|-----------------------|------------------------|---------------------------|----------------------------------|
| 0 | 116 | 1.67 ± 0.69 | 0.47 ± 0.63 | -1.03 ± 0.41 |
| 1 | 195 | 0.59 ± 0.39 | -0.87 ± 0.71 | 0.41 ± 0.64 |
| 2 | 202 | -0.25 ± 0.67 | -0.41 ± 0.64 | -0.99 ± 0.45 |
| 3 | 154 | -1.12 ± 0.64 | -0.77 ± 0.63 | 1.26 ± 0.60 |
| 4 | 274 | -0.31 ± 0.54 | 1.15 ± 0.75 | 0.16 ± 0.70 |

Cluster 0 has the highest number of impressions but the lowest engagement rate, indicating that viewers in this group watch a lot of videos but rarely interact.

1. Cluster 1 shows the opposite consumption pattern, with shorter video durations but fairly high engagement levels. Cluster 2 has a low number of impressions and engagement, reflecting that they are passive viewers.
2. Cluster 3 has a low number of impressions but a very high engagement rate, indicating that they tend to be active in interacting despite the lower number of videos watched.
3. Cluster 4 has the highest video duration with a medium number of views and medium engagement, indicating that this group enjoys long-form content more.

Distribution of Video Categories in Each Cluster

The distribution of video categories in each cluster table is shown in the following table:

Table 2. Distribution of Video Categories in Each Cluster

| Category | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|------------------|-----------|-----------|-----------|-----------|-----------|
| Film & Animation | 31 | 12 | 19 | 4 | 15 |

| Category | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|------------------|-----------|-----------|-----------|-----------|-----------|
| Autos & Vehicles | 0 | 0 | 1 | 0 | 4 |
| Music | 13 | 12 | 9 | 41 | 10 |
| Pets & Animals | 0 | 0 | 0 | 1 | 2 |
| Sports | 10 | 6 | 23 | 9 | 21 |
| Travel & Events | 0 | 1 | 1 | 2 | 2 |
| Gaming | 24 | 27 | 16 | 11 | 46 |
| People & Blogs | 4 | 56 | 17 | 14 | 41 |
| Comedy | 0 | 3 | 2 | 3 | 9 |
| Entertainment | 16 | 68 | 31 | 36 | 70 |
| News & Politics | 0 | 3 | 11 | 27 | 29 |
| Howto & Style | 8 | 1 | 29 | 2 | 14 |
| Education | 3 | 3 | 5 | 1 | 4 |
| Science & Tech | 7 | 3 | 38 | 3 | 7 |

Table 3. Distribution of Dominant Categories in Each Cluster

| Cluster | Dominant Video Categories | Number of Videos |
|---------|---|------------------|
| 0 | Film & Animation, Gaming, Entertainment | 31, 24, 16 |
| 1 | Entertainment, People & Blogs | 68, 56 |
| 2 | Science & Technology, Education, Sports | 38, 29, 23 |
| 3 | Music, Entertainment, News & Politics | 41, 36, 27 |
| 4 | Entertainment, Gaming, News & Politics | 70, 46, 29 |

1. Cluster 0 watches more videos in the movie, animation and gaming categories, which tend to have longer durations but less interaction.
2. Cluster 1 is dominated by the entertainment and vlog categories, which tend to be shorter and more interactive.
3. Cluster 2 consumes more educational, science, technology and sports content, with more targeted consumption patterns but low levels of interaction.
4. Cluster 3 shows an interest in music and news content, with a tendency to watch short-form videos but is highly engaged.
5. Cluster 4 enjoys various categories of entertainment, gaming and politics with a longer duration than the other clusters.

Cluster Interpretation

After clustering using K-Means, five groups of YouTube viewers in Indonesia with different characteristics were found. The analysis was conducted based on the number of views (Log_Views), duration of videos consumed (Log_Duration), and engagement rate (Engagement Rate). Each cluster is named based on its video consumption pattern and engagement rate.

1. Cluster 0: Passive Observer with High Consumption

Viewers in this cluster watch a lot of videos but have low engagement. They tend to be more passive, rarely commenting or interacting with the videos they watch. Medium-

length videos are the top choice, and the dominant category shows that they enjoy longer-form content such as animated movies, entertainment or games.

2. Cluster 1: Fast Consumer with Active Interaction

This cluster has an audience that is more active in interacting, such as commenting, liking or sharing videos. They prefer short videos that are quick to consume, such as vlogs or entertainment snippets. High engagement indicates a strong community bond between viewers and content creators in this category.

3. Cluster 2: Deep Audience with a Focus on Education

Viewers in this cluster prefer information-based content, such as educational and science videos. They watch for a relatively longer duration than other clusters, indicating a tendency to focus more on understanding a topic. However, the low level of engagement suggests that they are more passive consumers of information than active participants in discussions.

4. Cluster 3: Spontaneous Viewers with Short Entertainment Preferences

Viewers in this cluster prefer short videos and have a tendency to move from one video to another. They are also very active in responding to the videos they watch, such as liking and commenting. The dominance of the music and news categories suggests that they use YouTube as a means of quick entertainment or a source of quick information.

5. Cluster 4: Diverse Audience with Stable Consumption

This cluster has viewers with diverse content preferences, such as entertainment, politics, and gaming. They watch videos in large numbers and for longer durations. This indicates a tendency to enjoy in-depth discussion or documentary content. The moderate level of engagement shows that they are not as active as clusters 1 and 3 in interacting, but still have a tendency to respond to videos that catch their attention.

Content Strategy

Cluster 0 requires strategies that increase audience interaction. One way to do this is to add interactive elements to the video such as polls, challenges, or invitations to discussion in the comments section. In addition, creators can increase engagement by inserting storytelling elements or open-ended questions that invite responses from the audience.

In Cluster 1, a strategy that can be applied is to create a video with a concise format that captures attention within the first few seconds. The use of interactive elements such as viral challenges, behind-the-scenes vlogs, or collaborations with other content creators can also increase engagement.

In Cluster 2, creators can add more interesting visual elements, such as animations or interactive illustrations, to keep the audience's attention longer. In addition, engagement can be increased by including discussions or Q&A sessions to encourage more active interaction.

For Cluster 3, an effective strategy is to optimize short video formats such as YouTube Shorts or snippets that directly convey the point of the video. Using trending elements, such as popular music or current affairs, can also increase engagement.

In Cluster 4, content needs to be designed to keep the viewer's attention for a long duration. The use of features such as timestamps or video chapters can help improve

retention, especially for long-form videos. In addition, engaging storytelling and in-depth discussions will further strengthen the appeal of videos for this cluster.

CONCLUSION

The application of K-Means to segment YouTube viewers in Indonesia based on video consumption patterns, is able to produce segmentation with fairly good performance. The clustering results form five groups of viewers with different characteristics. Cluster 0 consists of viewers with a high number of views who like Gaming and Film & Animation themed videos but have a low level of engagement. Cluster 1 includes entertainment and vlog viewers with a tendency for more active interaction, while Cluster 2 is dominated by educational and technology video viewers with medium duration preferences but low engagement levels. Cluster 3 consists of music and news video viewers who consume short-form videos with high levels of engagement. Cluster 4 includes viewers with more diverse content preferences and longer viewing durations than the other clusters. These findings provide insights into the video consumption patterns of Indonesian YouTube viewers and the implications for content creators and digital industry players. Based on the analysis, effective content strategies can be tailored to the characteristics of each cluster, such as increasing interaction for passive viewers, optimizing short video formats, or using storytelling elements in long-form videos. This study still has some limitations, such as the limited number of features and the video categories analyzed only include video categories that are on the trending list, not all video categories on YouTube Indonesia. Future research can expand the scope of analysis by considering more variables, such as adding features related to audience demographics, expanding the video categories analyzed, and considering individual preferences or seasonal trends in content consumption. Additionally, other algorithmic approaches, such as DBSCAN or Hierarchical Clustering, can be compared to evaluate the effectiveness of more accurate segmentation.

REFERENCE

- [1] Wahidin, W., Mugihartadi, M., Aviani, T. H. B., Pratiwi, H., Wijaya, Y. I., Andie, A., Windarto, A. P., and Waluyo, A. 2021. Application of data mining techniques using the K-Means Method on Unmet Need of Health Services by Province in Indonesia. *Journal of Physics: Conference Series*, 1783 (1), p.012012. doi:10.1088/1742-6596/1783/1/012012.
- [2] Han, J., Kamber, M., and Pei, J. *Data mining: Concepts and techniques*. 4th ed. Amsterdam: Elsevier, 2022.
- [3] Jain, A. K. *Data clustering: Theory, algorithms, and applications*. 1st ed. Cham: Springer, 2023.
- [4] DataReportal. *Essential YouTube Stats: Everything You Need to Know*. (Updated Jan 2025). Available at: <https://datareportal.com/essential-youtube-stats> [Accessed Apr 4, 2025].
- [5] Ipsos, "Survei: YouTube layanan video paling disukai Gen Z di Indonesia," *Tempo*, vol. 39, no. 2, pp. 45–57, 2023. [Online]. Available: <https://www.tempo.co/digital/survei->

- youtube-layanan-video-paling-disukai-gen-z-di-indonesia-133226. [Accessed Apr. 4, 2025].
- [6] Hootsuite and We Are Social, "Indonesian digital report 2023," *DataReportal*, Jan. 18, 2023. Available: <https://datareportal.com/reports/digital-2023-indonesia>. [Accessed Apr. 4, 2025].
- [7] Rahman and A. Nugroho, "Segmentasi pengguna YouTube berdasarkan perilaku interaksi menggunakan K-Means clustering," *J. Teknol. Inf. dan Komun.*, vol. 10, no. 1, pp. 55–68, 2022.
- [8] S. Sitanggang, F. Umbara, and H. Ashaury, "Klasifikasi video pada media sosial YouTube dengan menggunakan metode K-Means dan Support Vector Machine," *J. Locus Penelitian dan Pengabdian*, vol. 2, pp. 1027–1032, Nov. 2023, doi: 10.58344/locus.v2i10.1732.
- [9] K. Widjaja and R. Oetama, "K-Means clustering video trending di YouTube Amerika Serikat," *Ultima InfoSys: J. Ilmu Sistem Informasi*, vol. 11, no. 2, pp. 78–84, 2020, doi: 10.31937/si.v11i2.1508.
- [10] S. A. Perdana, S. F. Florentin, and A. Santoso, "Analisis segmentasi pelanggan menggunakan K-Means clustering studi kasus aplikasi Alfagift," *Sebatik*, vol. 26, no. 2, pp. 446–457, 2022, doi: 10.46984/sebatik.v26i2.1991.
- [11] S. Salmon, A. Azahari, and H. Ekawati, "Perbandingan kinerja algoritma K-Nearest Neighbor dan algoritma Random Forest untuk klasifikasi data mining pada penyakit gagal ginjal," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, pp. 1943–1953, 2024, doi: 10.47065/bits.v6i3.6476.