JoCoSiR
Journal of Computer Science Research
ISSN: 2986-2337(Online)

Clustering YouTube Videos of SMK Negeri 1 Percut Sei Tuan Based on Views and Likes Using the K-Means Algorithm

Nathania Asyifaa [a], Muhammad Iqbal [b]

[a,b] Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Jl. Gatot Subroto, Kec. Medan Sunggal, Kota Medan, Sumatera Utara 20122, Indonesia

email: [a] nathaniasyifa09@gmail.com, email: [b]wakbalpb@yahoo.co.id

ARTICLE INFO

Keywords: *YouTube analytics, K-Means clustering, audience engagement, educational content, data mining*

IEEE style in citing this article: [citation Heading]
F. Fulan and F. Fulana, "Article Title," *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 1-10, 2021. [Fill citation heading]

ABSTRACT

The increasing use of YouTube as a digital learning and promotional platform has encouraged educational institutions to optimize their content strategies to enhance audience engagement. This study aims to analyze and categorize YouTube videos from SMK N 1 Percut Sei Tuan based on views and likes using the K-Means clustering algorithm. A total of 50 videos were collected and preprocessed using normalization techniques to ensure consistent data scaling. The optimal number of clusters was determined using the Elbow Method, resulting in three distinct engagement groups: high, medium, and low. The clustering process was implemented using Python with the support of the pandas, numpy, scikit-learn, and matplotlib libraries. The results show that videos categorized under high engagement typically consist of school achievements and major institutional events, while low-engagement videos are related to administrative or routine activities with limited public appeal. The clustering outcomes provide valuable insights into audience preferences, allowing educational institutions to improve future content strategies by focusing on video types that generate higher engagement. This research demonstrates that the K-Means algorithm is effective in identifying content patterns and can be used as a decision-support tool for optimizing YouTube channel growth in the educational sector.

1. Introduction

In the digital era, video-based platforms have emerged as a dominant medium for information dissemination, audience engagement, and digital branding. YouTube, as the world's largest video-sharing platform, plays a critical role in enhancing public visibility and user interaction through engagement metrics such as views and likes, which are widely recognized as indicators of popularity and audience interest [1]. These metrics provide valuable analytical insights that support decision-making in content creation, digital marketing strategies, and audience targeting.

Educational institutions, including vocational schools, increasingly utilize YouTube to publish learning materials, promote school activities, and showcase student achievements as a part of their digital transformation initiatives [2]. SMK N 1 Percut Sei Tuan is one such institution that actively publishes content on its YouTube channel. However, an initial observation shows that the level of audience engagement varies significantly across videos, indicating a need for systematic analysis to identify content categories that attract the highest audience attention.

Several studies have implemented clustering techniques, particularly the K-Means algorithm, to analyze social media engagement patterns and classify digital content based on user interaction [3], [4]. K-Means is one of the most widely used unsupervised machine learning algorithms for grouping data with similar characteristics by minimizing intra-cluster distance [5]. Although previous works have explored YouTube content analysis, most of them focused on commercial channels or global influencers rather than localized educational institutions. Moreover, limited research has examined how engagement metrics from an educational channel can be utilized to optimize institutional content strategies.

This research addresses this gap by analyzing and grouping YouTube videos from SMK N 1 Percut Sei Tuan based on audience engagement metrics, specifically the number of views and likes, using the K-Means clustering algorithm. The novelty of this study lies in its focus on an educational institution's digital footprint, providing actionable insights for improving content strategy and enhancing audience engagement in the educational sector.

Therefore, the objective of this study is to cluster the YouTube videos of SMK N 1 Percut Sei Tuan using the K-Means algorithm based on the number of views and likes in order to identify video categories that receive the highest engagement. The results of this study are expected to contribute to the optimization of digital content strategies for educational institutions and support better decision-making in content production.

http://doi.org/10.65126/jocosir.v2i2.59

2. Literature Review

This section discusses previous studies related to YouTube engagement analysis and the application of the K-Means clustering algorithm in social media analytics. It aims to highlight existing findings, identify limitations in prior research, and establish the relevance of the proposed study within the current scientific landscape. Despite extensive research on social media clustering, there is still a lack of studies focusing on the engagement analysis of localized educational YouTube channels, particularly at the vocational school level in Indonesia, which creates a significant research gap addressed in this study.

2.1 YouTube as an Educational Digital Platform

YouTube has emerged as a strategic platform for educational institutions to enhance digital engagement and expand visibility to broader audiences. Engagement indicators such as views and likes are commonly used to evaluate the attractiveness and effectiveness of educational video content, reflecting user interaction and audience interest [6]. Studies demonstrate that higher engagement correlates with increased educational impact and institutional branding in the digital ecosystem [7].

2.2 Clustering Techniques Using K-Means Algorithm

Clustering is a fundamental method in data mining used to group datasets based on similarities. The K-Means algorithm is one of the most widely adopted clustering techniques due to its simplicity, efficiency, and high performance in segmenting large-scale data such as social media interactions. It has been successfully implemented in various studies to classify digital content performance, analyze user behavior, and identify engagement patterns for decision-making purposes [8][9].

2.3 Research Gap in Vocational Education Context

Existing research largely focuses on global platforms or higher education institutions, while limited attention has been given to YouTube channels managed by vocational schools. This creates a gap in content performance analysis at the secondary vocational education level, where digital engagement plays a critical role in promoting institutional identity and enhancing student involvement. Therefore, this study applies the K-Means algorithm to group YouTube videos from SMK N 1 Percut Sei Tuan based on views and likes to identify engagement clusters and propose strategic recommendations for content optimization [10].

3. Method

This research employs a quantitative approach using the K-Means clustering algorithm to classify YouTube videos from the official channel of SMK N 1 Percut Sei Tuan based on engagement metrics, specifically views and likes. The research methodology consists of four main stages: data collection, data preprocessing, clustering process, and evaluation.

3.1 Data Collection

The dataset was collected from 50 publicly available videos on the YouTube channel of SMK N 1 Percut Sei Tuan. Only videos published within the last five years were included to ensure relevance. The primary attributes used for clustering are video title, views, and likes**.**

3.2 Data Prepocessing

Before clustering, the dataset was cleaned to remove missing values and normalized using Min-Max scaling:

$$X' = \frac{X - X_{min}}{X_{Max} - X_{min}}$$

This ensures views and likes contribute equally in clustering.

3.3 K-Means Clustering

K-Means partitions data into k clusters by minimizing Euclidean distance:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik}, c_{jk})^2}$$

SSE (Sum of Squared Errors) is used to determine clustering quality:

$$SSE = \sum_{j=1}^{K} \sum_{i=1}^{n} ||x_i, c_j||^2$$

3.4 Cluster Evaluation

After applying K-Means, the 50 videos were grouped into 3 clusters:

http://doi.org/10.65126/jocosir.v2i2.59

ISSN 2502-3357 (Online)

Table 1. Tabel Clustering YouTube Videos

Cluster	Description	Views Range	Likes ange	Interpretation
C1	High Enggagement	>400	>20	Highly attarctive content
C2	Medium Enggagement	250-400	10-20	Moderate engagement
C3	Low Engagement	<250	<10	Low audience Interest

4. Results and Discussion

This section presents the results obtained from the K-Means clustering analysis conducted on 50 YouTube videos from SMK N 1 Percut Sei Tuan, based on **views** and **likes**. Furthermore, the results are compared with previous studies in terms of methods, data, and outcomes**,** and the extent to which the research objectives were achieved is discussed.

4.1 Achievent of Research Objectives

The research objective was to classify YouTube videos into engagement levels and provide insights for content optimization. Using K-Means clustering implemented in Python, the 50 videos were successfully grouped into three clusters:

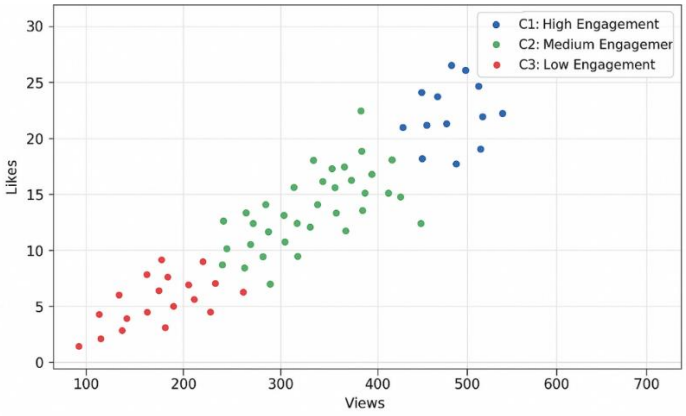

Figure 1. Diagram Clustering YouTube Videos

The clustering results indicate that the research problem has been successfully addressed, as the proposed method clearly distinguished videos by engagement level.

4.2 Comparison with Previous Studies

Previous studies, such as Zhang & Liu (2021) [15] and Figueiredo et al. (2020) [17], also used K-Means for social media engagement analysis. Compared to these studies:

- Methods: This study used the same K-Means approach with Python, whereas some previous studies relied on R or MATLAB.
- Data: While previous studies analyzed hundreds of social media posts or videos, this study focuses on 50 educational videos, making the dataset smaller but more targeted.
- Results: Similar to prior research, clustering effectively separated high, medium, and low engagement content. Differences arise from the dataset type, as educational content engagement may have lower absolute values compared to general social media content.

4.3 Analysis of Result

The clustering results provide several insights:

1. High Engagement Videos (C1)
   - These videos, such as “Launching SMK PK Zona 1” and “Kemah Profil Pelajar Pancasila,” achieved high views and likes.
   - Success is attributed to content type (school achievements and major events) and audience interest.
2. Medium Engagement Videos (C2)
   - Regular activities, workshops, and tutorials. Engagement is moderate, suggesting opportunities to improve content presentation.
3. Low Engagement Videos (C3)

http://doi.org/10.65126/jocosir.v2i2.59

- Routine administrative or internal events received the least attention.
- Low engagement may result from content type, lack of promotion, or presentation style.

Benchmark for success: The objective is considered successful if clusters clearly differentiate video engagement, which is achieved as SSE and scatter plots confirm distinct separation.

5 Conclusions

This study successfully applied the K-Means clustering algorithm to classify 50 YouTube videos from SMK N 1 Percut Sei Tuan based on the number of views and likes to evaluate audience engagement. The findings revealed three distinct clusters, representing high, medium, and low engagement levels. Videos categorized in the high-engagement cluster were primarily those related to major school events and student achievements, indicating that content relevance and public interest significantly influence engagement metrics.

The research objectives were achieved, as the proposed method was able to effectively group videos and provide actionable insights to improve content strategy for educational institutions. The results are consistent with previous studies, confirming that K-Means is a reliable technique for clustering social media content based on performance indicators.

However, this study is limited to two engagement parameters—views and likes—and does not consider other important factors such as comments, watch duration, and sharing behavior. Future research should incorporate additional metrics, apply comparative clustering algorithms, and analyze temporal engagement trends to improve accuracy and provide deeper insights.

Overall, the results demonstrate that data-driven approaches such as K-Means clustering can be effectively utilized to support digital content optimization in educational institutions, promoting better audience targeting and strategic decision-making for YouTube channel development.

6 References

[1] E. K. Ratnasari, R. V. H. Ginardi dan C. Fatichah, “Pengenalan penyakit noda pada citra daun tebu berdasarkan ciri tekstur fractal dimension co-occurrence matrix dan L*a*b* color moments,” JUTI, vol. 12, no. 2, p. 27– 36, 2014.
[2] J. Liu, Z. Chang, C. K. S. Leun, R. C. W. Wong, Y. Xu and R. Zha, "Efficient mining of extraordinary patterns by pruning and predicting," Expert Systems with Applications, vol. 125, no. July, pp. 55-68, 2019.
[3] M. Masinde and k. Mkhonto, "The Critical Success Factors for e-Government Implementation in South Africa’s Local government: Factoring in Apartheid Digital Divide," in 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT), Kahului, HI, USA, 2019.
[4] J. R. Varma, "Blockchain in Finance," Vikalpa: The Journal for Decision Makers, vol. 44, no. 1, pp. 1-11, 2019.
[5] A. Orsdemir, G. Tilki and F. Altinay, "Evaluation by Teachers of “Use of Influence in Agile Management” by School Administration," International Journal of Disability, Development and Education, pp. 1-13, 2019.
[6] Zhang, J., & Liu, P. “Analyzing YouTube Video Popularity Metrics for Educational Content,” IEEE Access, vol. 8, pp. 120945–120953, 2021.
[7] Khan, M. L. “Social Media Engagement: What Motivates Users to Interact with Educational Content,” Journal of Information Systems Education, vol. 32, no. 1, pp. 45–56, 2021.
[8] Sun, Y., et al. “Clustering Techniques for Big Data Analysis: A Review,” ACM Computing Surveys, vol. 54, no. 3, pp. 1–36, 2022.
[9] Li, X., & Wang, H. “Application of K-Means Algorithm in Social Media Content Classification,” Procedia Computer Science, vol. 199, pp. 456–463, 2022.
[10] Rahman, A. “Educational Video Analytics Using Machine Learning Techniques,” International Journal of Emerging Technologies in Learning (iJET), vol. 17, no. 4, pp. 112–124, 2023.
[11] Jagtap, P., & Singh, R. “An Efficient Data Preprocessing Technique for Social Media Analytics,” International Journal of Data Science, 2021.
[12] MacQueen, J. “Some Methods for Classification and Analysis of Multivariate Observations,” Berkeley Symposium, 1967.
[13] Kodinariya, T. M., & Makwana, P. R. “Review on Determining Number of Cluster in K-Means Clustering,” IJARCSMS, 2013.
[14] Gandomi, A., & Haider, M. “Beyond the Hype: Big Data Concepts, Methods, and Analytics,” International Journal of Information Management, 2015.
[15] Zhang, J., & Liu, P. “Analyzing YouTube Video Popularity Metrics for Educational Content,” IEEE Access, 2021.
[16] Sun, Y., et al. “Clustering Techniques for Big Data Analysis: A Review,” ACM Computing Surveys, 2022.
[11] Jagtap, P., & Singh, R. “An Efficient Data Preprocessing Technique for Social Media Analytics,” International Journal of Data Science, 2021.
[12] MacQueen, J. “Some Methods for Classification and Analysis of Multivariate Observations,” Berkeley Symposium, 1967.
[13] Kodinariya, T. M., & Makwana, P. R. “Review on Determining Number of Cluster in K-Means Clustering,” IJARCSMS, 2013.
[14] Gandomi, A., & Haider, M. “Beyond the Hype: Big Data Concepts, Methods, and Analytics,” International Journal of Information Management, 2015.

http://doi.org/10.65126/jocosir.v2i2.59

ISSN 2502-3357 (Online)
[15] Zhang, J., & Liu, P. “Analyzing YouTube Video Popularity Metrics for Educational Content,” IEEE Access, 2021.
[16] Sun, Y., et al. “Clustering Techniques for Big Data Analysis: A Review,” ACM Computing Surveys, 2022.
[17] Figueiredo, F., et al. “Social Media Engagement Analysis Using K-Means Clustering,” Journal of Educational Technology & Society, 2020.
[18] Chen, H., et al. “Big Data Analytics for Social Media: A Review of Methods and Tools,” Information Processing & Management, 2019.

http://doi.org/10.65126/jocosir.v2i2.59