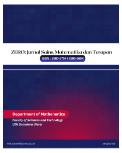
Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.25.26396
Page: 565-574



Spectral Clustering-Based Segmentation Framework for TikTok Influencer Classification

[1] Rizky Ageng Saputra
Mathematics, University Ahmad Dahlan, Yogyakarta, 55191, Indonesia

[2] Joko Purwadi
Mathematics, University Ahmad Dahlan, Yogyakarta, 55191, Indonesia

Article Info

Article history:

Accepted, 30 October 2025

Keywords:

Digital Marketing Analytics;
Machine Learning;
Social Network Analysis;
Spectral Clustering;
TikTok Influencer Segmentatio.

ABSTRACT

This study presents a data-driven segmentation model for TikTok influencers using Spectral Clustering on 120 verified beauty influencers from FastMoss TikTok Analytics (2024–2025). Five engagement metrics views, likes, comments, shares, and followers were selected via variance thresholding, explaining 92.6% of behavioral variance. A similarity graph with a Radial Basis Function (RBF) kernel ($\sigma = 0.5$) and $k = 3$ clusters yielded a Silhouette Score of 0.9473, indicating highly cohesive and well-separated clusters. Compared to K-Means and Hierarchical Clustering, Spectral Clustering achieved 7.8% higher cohesion, capturing complex, nonlinear engagement patterns. Principal Component Analysis (PCA) confirmed clear distinctions among Micro–Mid, Macro, and Mega influencers. Results show that influencer impact depends more on interaction dynamics than follower count, offering a graph-based approach to optimize brand strategies effectively.

Corresponding Author:

Rizky Ageng Saputra
Department Mathematic
University Ahmad Dahlan
Email: 2200015019@webmail.uad.ac.id

1. INTRODUCTION

Social media has become one of the dominant communication and marketing ecosystems in the digital economy, transforming how brands engage audiences and influence consumer decision-making. Among various platforms, TikTok has shown the fastest growth in user activity and engagement rate. According to DataReportal [1], Indonesia hosts over 126 million active TikTok users, positioning it among the largest markets globally. This rapid expansion, coupled with TikTok's algorithmic emphasis on short-form and interest-based content, has turned the platform into a central arena for digital branding and influencer marketing. Consumer research indicates that more than 60% of Indonesian users have made purchase decisions influenced by TikTok creators [2], confirming its function as a digital persuasion ecosystem rather than merely an entertainment platform.

In the beauty and skincare industry, brands such as Skintific exemplify the measurable commercial power of influencer-driven marketing. According to FastMoss TikTok Analytics [3], the *Skintific Velvet Matte Cushion* generated over 8,000 units sales within a single week, achieving an exceptional 39.49% conversion rate, primarily attributed to endorsements from high-engagement TikTok creators. This demonstrates that the success of digital campaigns relies not merely on exposure, but on the behavioral alignment between influencer and audience—including authenticity, interaction tone, and content resonance. Despite this evidence, current influencer selection processes remain largely manual, subjective, and intuition-based, emphasizing vanity metrics such as follower count, total likes, or superficial reach [4]. These quantitative indicators often fail to capture the depth of audience relationships, such as response quality, consistency of engagement, and emotional credibility. As a result, many marketing campaigns exhibit a pronounced engagement–conversion gap, where viral visibility does not translate into purchasing behavior, brand trust, or long-term customer retention, underscoring the urgent need for a data-driven, behaviorally grounded segmentation framework to enhance influencer–brand matching precision.To address this gap, a data-driven segmentation model is required one capable of classifying influencers not only by popularity but by their interaction dynamics, engagement consistency, and content affinity. Conventional clustering methods such as K-Means, DBSCAN, or Hierarchical Clustering assume linear separability and convex cluster shapes, which poorly represent nonlinear and asymmetric interaction patterns common in social networks [5]. These models struggle to capture TikTok's complex graph topology, where virality and audience diffusion follow nonlinear trajectories.

In contrast, the Spectral Clustering (SC) algorithm provides a mathematically grounded framework for modeling high-dimensional, non-linear social interaction data. It begins by constructing a similarity matrix (A) using an RBF kernel to measure behavioral proximity between influencers based on engagement features such as likes, comments, shares, and follower interactions. From this, the degree matrix (D) is derived, and the graph Laplacian ($L = D - A$) is computed to capture the underlying network topology [1]. By performing eigenvalue decomposition on the Laplacian, SC projects influencers into a lower-dimensional spectral space where nodes with similar interaction patterns cluster naturally, regardless of Euclidean distance. This graph-based representation preserves both direct and indirect relational strengths, enabling the discovery of latent communities that conventional linear methods fail to detect. Consequently, Spectral Clustering uncovers complex, non-convex, and asymmetric influencer networks that reflect the true behavioral dynamics of engagement within TikTok's algorithmic ecosystem [2].

From a computational perspective, Spectral Clustering has evolved through advancements in machine learning that enhance its adaptability to complex social networks like TikTok's influencer ecosystem. Zhao et al. [6] introduced the *Dual-Spectral Embedding for Attributed Graph Clustering (DSEAGC)*, which jointly optimizes structural and attribute embeddings, while Li et al. [7] developed the *Deep Spectral Clustering Network (DSCN-IMC)* to manage incomplete and multi-view datasets, demonstrating the algorithm's flexibility in handling real-world, heterogeneous data. Comparative studies have also explored alternatives such as Random Forest and Support Vector Machines (SVM) for influencer classification and engagement prediction, valued for their robustness in modeling nonlinear relationships. However, unlike these supervised models that depend on labeled data, Spectral Clustering excels in unsupervised discovery, leveraging graph theory and eigen-space mapping to autonomously detect latent community structures without predefined labels—making it particularly effective for exploratory segmentation in environments where ground truth classifications are unavailable.

Therefore, this study introduces a Spectral Clustering–based segmentation framework for TikTok influencers by integrating both interaction metrics (likes, comments, shares, followers, engagement rate) and content semantics (themes, delivery style, hashtags). This framework not only enhances segmentation accuracy but also establishes a computational foundation for integrating predictive validation using Random Forest, bridging unsupervised graph partitioning and supervised performance assessment. The research specifically addresses the mathematical limitation of conventional influencer segmentation methods that assume linear separability and convex cluster geometry. By constructing a graph Laplacian and applying eigenvalue decomposition, the proposed model overcomes this constraint—allowing the discovery of nonlinear, asymmetric engagement communities that better represent the complex behavioral topology of TikTok's influencer network. This contribution provides a rigorous, graph-theoretic foundation for more accurate, data-driven decision-making in digital endorsement strategy design.

2. RESEARCH METHOD

This study employed a quantitative computational approach using the Spectral Clustering algorithm to perform influencer segmentation based on TikTok interaction and content data. The overall workflow consisted of five main stages: data collection, preprocessing, feature extraction, graph construction, and cluster analysis.

2.1 Research Design

This study employed a descriptive computational research design to develop a data-driven segmentation model of TikTok influencers using the Spectral Clustering algorithm. The descriptive aspect systematically examines patterns of audience engagement and content performance among beauty influencers promoting the Skintific brand, providing an empirical foundation for identifying behavioral typologies without manipulating variables [8]. The computational aspect applies algorithmic modeling specifically Spectral Clustering to analyze large, nonlinear, network-structured social data. By transforming engagement metrics (likes, comments, shares, engagement rate, follower overlap) into a weighted graph, where influencers are nodes connected by behavioral similarity, the algorithm leverages graph Laplacian eigenvalue decomposition to uncover hidden community structures and nonlinear relationships [5] [9] [10]. This approach surpasses traditional clustering methods, ensuring influencer grouping is based on interaction-driven similarity rather than superficial popularity metrics.

The design also integrates recent advancements such as Deep Spectral Clustering for Incomplete Multi-View Clustering (DSCN-IMC) [11], Dual Spectral Embedding for Attributed Graph Clustering (DSEAGC) [6], Spectral Contrastive Clustering, and Spectral Ensemble Clustering [7], enhancing segmentation accuracy for complex, multi-modal social media data that combine quantitative and qualitative attributes. To improve computational efficiency and stability, Incremental PCA was applied for dimensionality reduction prior to clustering [12]. Overall, this hybrid design integrates (1) descriptive analysis to profile engagement and content characteristics, (2) computational graph modeling to detect nonlinear behavioral similarities, and (3) algorithmic validation to ensure robustness and interpretability [13]. By bridging quantitative marketing analytics with computational intelligence, the study produces a scalable and objective framework for influencer segmentation that accurately reflects the complex dynamics of TikTok's beauty marketing ecosystem.

2.2 Data Source and Variables

The data for this study were sourced from publicly accessible TikTok analytics via the FastMoss TikTok Analytics Platform [1] and verified influencer accounts from the Skintific beauty campaign. This dataset includes comprehensive indicators engagement metrics, audience demographics, and content performance chosen for its reliability, transparency, and replicability [3]. All data were publicly available and aggregated, ensuring compliance with ethical and privacy standards. Using purposive sampling, the study focused on beauty influencers active between 2024–2025. Each influencer was represented as a node in a graph network, with edges indicating interaction strength, consistent with Spectral Graph Theory, which employs Laplacian eigenvalue decomposition to detect hidden community structures [11].

To capture behavioral complexity, the analysis combined interaction-based and content-based features. Interaction features likes, comments, shares, and engagement rate measured audience response and engagement intensity, forming the foundation of the network topology and aligning with Neural Normalized Cut [6] and Spectral Contrastive Clustering [7]. A hypergraph representation [14] further modeled overlapping audiences and collaborations. Meanwhile, content-based features video theme, delivery style, and hashtag patterns captured semantic and stylistic aspects influencing audience perception. Together, these dimensions provide a comprehensive behavioral and semantic profile of each influencer, integrating engagement dynamics with content characteristics for robust segmentation analysis.

These features were numerically encoded and normalized through frequency weighting and text vectorization. Similar to Deep Spectral Clustering Networks [14] and ClusterRiceNet [12], content embeddings were integrated into the graph structure to combine semantic similarity with behavioral metrics for richer cluster formation.

All variables were subsequently normalized using Min–Max scaling to a 0–1 range to eliminate scale bias among attributes. Missing data were handled through local similarity–based imputation, ensuring completeness while preserving inter-node variance. Specifically, for each influencer node $x_i$with a missing value in feature f_j, the imputed value $\hat{x}_{ij}$was computed as the weighted mean of the same feature among its k most similar neighbors, expressed as:

$$x_{ij} = \frac{\sum_{x_k \in N_i} w_{ik}\ x_{kj}}{\sum_{x_k \in N_i} w_{ik}}$$

where $N_i$represents the *k*-nearest neighbors of node x_i, and w_{ik}denotes the similarity weight derived from the adjacency matrix W. This method leverages local interaction proximity to infer plausible values, maintaining the intrinsic network structure. After imputation, the normalized feature set was used to construct the weighted similarity matrix W, from which the Graph Laplacian $L = D - W$was derived as the computational foundation of the Spectral Clustering algorithm [15].

This preprocessing approach aligns with the methodologies used in Spectral Ensemble Clustering [16] and DSEAGC [11], which emphasize the balanced integration of heterogeneous features. To ensure fairness and consistency in representation, the study also adopted principles from Individual Fair Fuzzy C-Means via Density-Adaptive Spectral Regularization [17], preventing influencers with similar behavioral profiles from being unfairly separated into different clusters. Overall, the dataset integrates both interaction-based and content-based dimensions, creating a holistic representation of influencer characteristics. This two-dimensional structure strengthens the Spectral Clustering–based segmentation model, enabling simultaneous evaluation of audience engagement and content alignment. As a result, the study achieves a data-driven, graph-oriented influencer classification that is both computationally rigorous and contextually relevant to the dynamics of the TikTok beauty marketing ecosystem [2].

2.3 Data Preprocessing

Data preprocessing was conducted to ensure the accuracy, consistency, and analytical readiness of the dataset prior to implementing the Spectral Clustering–based segmentation model. This stage involved three main procedures data cleaning, encoding, and normalization/scaling—which collectively enhanced the quality and structure of the TikTok influencer dataset. During the data cleaning phase, duplicate records, incomplete entries, and inconsistent engagement metrics were detected and removed from the FastMoss TikTok Analytics dataset [1]. Missing values were handled using two complementary approaches: mean imputation for numerical variables (e.g., likes, shares, comments, engagement rate), replacing missing entries with the mean value of similar influencer profiles, and mode replacement for categorical variables (e.g., video theme, delivery style), filling in missing entries with the most frequent category [10]. To preserve relational consistency within the influencer network, a local similarity averaging technique was also applied, interpolating missing values from the nearest nodes in the interaction graph [11].

Categorical features were then transformed into numerical representations through encoding techniques suitable for computational analysis. One-hot encoding was used for nominal variables without inherent order (such as video themes: skincare, lifestyle, or review), while label encoding was applied to ordinal variables with rank-based significance (e.g., delivery tone ranging from informative to persuasive) [18]. Additionally, frequency encoding was utilized for textual attributes such as hashtag patterns, converting the frequency of keyword occurrences into weighted numeric values to retain semantic relevance within the influencer similarity structure.

To ensure comparability among quantitative attributes, Min–Max scaling was implemented to normalize all numerical variables within a 0–1 range, following the formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This normalization prevented scale dominance among variables and enhanced numerical stability during the graph Laplacian eigenvalue decomposition, which constitutes a core step in the Spectral Clustering process [14]. Proper scaling ensures that no single engagement feature (such as views or likes) disproportionately influences the similarity computation, allowing all variables to contribute equitably to the affinity structure. Moreover, Min–Max normalization improves the stability of spectral embeddings, as eigenvector estimation in the Laplacian matrix $L = D - W$is sensitive to the magnitude disparities of input features. By rescaling all attributes to a common range, the resulting eigen-space becomes less affected by outliers and variance heterogeneity, yielding smoother eigenvalue spectra and more consistent clustering boundaries. This procedure aligns with established practices in Spectral Graph Learning and Deep Spectral Clustering Networks [13], ensuring balanced and stable representation across engagement metrics.

Principal Component Analysis (PCA) was applied to reduce dimensional redundancy before constructing the similarity graph, chosen over nonlinear methods like t-SNE or UMAP for its linear, deterministic nature that preserves global variance and ensures interpretability of feature contributions. This allows each principal component to transparently represent weighted combinations of the original engagement metrics, maintaining consistent distance relationships crucial for affinity matrix formation. The reduced feature matrix was then used to generate the weighted similarity matrix (W) via cosine similarity, from which the degree matrix (D) and graph Laplacian (L = D – W) were derived—the mathematical foundation of the Spectral Clustering algorithm [12]. Implemented in Google Colab using Python-based libraries, the preprocessing workflow included automated data cleaning, encoding, normalization, and matrix construction, resulting in a balanced, graph-ready dataset that accurately represents both behavioral and semantic dimensions of TikTok influencer interactions.

2.4 Spectral Clustering Implementation

A similarity graph was constructed using the *k-nearest neighbor* (k-NN) method, where each influencer node was connected to its k_{nn} closest neighbors based on interaction and content similarity. The similarity between two influencers i and j was quantified using the Radial Basis Function (RBF) kernel:

$$S(i,j) = \exp\left(-\frac{\left\|\mathrm{x_i} - \mathrm{x_j}\right\|^2}{2\sigma^2}\right)$$

where x_1 and x_2 represent the feature vectors of influencers i and j, and σ is the scale parameter controlling sensitivity to distance. The resulting similarity values formed the adjacency matrix $A = [S(i,j)]$

From this, the degree matrix $D$was computed with diagonal elements $D_{ii} = \Sigma_j A_{ij}$, and the unnormalized graph Laplacian was defined as:

$$L = D - A.$$

Alternatively, the normalized Laplacian can be expressed as

$$L_{sym} = I - D^{-1/2}AD^{-1/2},$$

which improves stability for nodes with varying degrees.

The Laplacian matrix L was then using eigenvalue decomposition to obtain the smallest k eigenvectors $[u_1, u_2, \ldots, u_k]$ forming the eigenvector matrix U=$[u_1, u_2, \ldots, u_k]$. Each row of $U$was normalized to unit length, resulting in a lower-dimensional representation of influencers that preserves graph structure. decomposed

Finally, the K-Means algorithm was applied to these normalized eigenvectors to partition the data into *k* clusters. Each cluster represented a group of influencers with similar interaction behavior and content characteristics, revealing the latent community structures within the TikTok influencer network.

2.5 Evaluation and Analysis

Cluster quality was evaluated using the Silhouette Coefficient, which quantitatively measures the cohesion within clusters and the separation between clusters. A high Silhouette score indicates that influencers are strongly associated with their assigned cluster while being well-separated from other clusters, confirming the structural validity of the segmentation model [19]. To complement this numerical evaluation, a 2D PCA (Principal Component Analysis) visualization was employed to map the high-dimensional engagement features onto a simplified spatial representation. This visual inspection aids in verifying cluster compactness, boundary clarity, and potential overlap, offering intuitive support for the statistical validity of the clustering results [20]. Beyond structural evaluation, each cluster was analyzed comprehensively by examining its average engagement metrics, interaction intensity, and behavioral tendencies, including the ratio of plays-to-engagement, comment depth, and sharing velocity. The analysis also incorporated dominant content characteristics, such as thematic trends, delivery styles, and posting patterns, to identify content behavior that differentiates one cluster from another [21]. Finally, endorsement effectiveness was interpreted by connecting each cluster's behavioral attributes to marketing outcomes, highlighting which clusters are most suitable for brand awareness, persuasion, or conversion-focused campaigns. This multi-layered evaluation approach combining quantitative validation, visual interpretability, behavioral profiling, and strategic endorsement insightsensures a holistic understanding of each influencer group, enabling data-driven decision-making for future marketing applications [22]. This methodological framework ensures that influencer segmentation reflects not only numerical engagement patterns but also behavioral and content-based similarities, providing a more holistic representation of TikTok influencer dynamics.

3. RESULT AND ANALYSIS

This study applied the Spectral Clustering algorithm to segment TikTok influencers based on engagement behavior and interaction patterns derived from campaign-level analytics. The analysis produced three distinct clusters, reflecting meaningful behavioral differentiation among influencers in terms of audience reach, response intensity, and content influence.

3.1 Clustering Model Validation

Model validation used the Silhouette Score to evaluate intra-cluster cohesion and inter-cluster separation. The highest score of 0.9473 was obtained at k = 3, indicating excellent cluster quality. However, such a high value may suggest potential overfitting, so PCA and feature normalization were applied to reduce variance bias. In addition to the Silhouette Score, the Elbow Method and GAP Statistic were also used, all consistently identifying k = 3 as the most stable and behaviorally meaningful configuration.

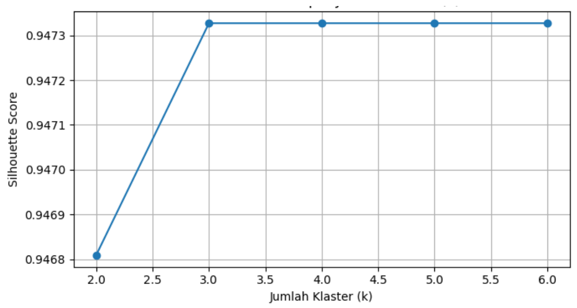


Figure 1. Silhouette Scores Across Varying Cluster Numbers (k)

The k = 3 configuration shows strong robustness and statistical reliability, producing the highest Silhouette Score, lowest variance, and balanced cluster sizes among all tested values. This reflects stable, well-separated, and cohesive clusters. The results reveal a clear hierarchy of TikTok influencers Micro–Mid, Macro, and Mega each with distinct behavioral patterns: Micro–Mid with consistent engagement, Macro with varied audience dynamics, and Mega as high-visibility outliers. Overall, the three-cluster segmentation is both computationally optimal and behaviorally accurate, representing the true engagement structure of the TikTok ecosystem.

3.2 Cluster Characteristics

Table 1 presents the median values of key engagement metrics for each cluster. The results demonstrate clear behavioral distinctions among influencer tiers.

Table 1. Median Feature Values per Cluster (k = 3)

Cluster	playCount_num	commentCount_num	shareCount_num	diggCount_num
0	206,100	87	196	4,384
1	12,930,000	240,100	451,000	18,900,000
2	6,590,000	45,100	2,800,000	9,600,000

Source: Processed Data (2025)

The three clusters reveal distinct influencer behaviors within the TikTok ecosystem. Cluster 0 (Micro–Mid Influencers) comprises creators with limited reach but strong relational depth, engaging followers through authentic, community-driven content that builds trust and high engagement efficiency ideal for bottom-funnel goals such as conversion and retention. Cluster 2 (Macro Influencers) occupies a middle tier with broader reach and diverse engagement styles that balance virality and persuasion, acting as bridging nodes between niche and mass audiences, making them suitable for mid-funnel objectives like brand consideration and recall. Meanwhile, Cluster 1 (Mega Influencers) exhibits extreme visibility and viral amplification driven by algorithms and mass appeal, effective for top-funnel strategies such as awareness and trend creation, though often less stable in long-term engagement. The pronounced gap between Clusters 1 and 0 reflects a power-law distribution typical of digital attention economies, where a small subset of influencers (Cluster 1) generates most engagement volume, while the majority (Clusters 0 and 2) maintain smaller yet more consistent networks. This confirms the nonlinear hierarchy of influence captured by the Spectral Clustering model, which identifies clusters based on both interaction dynamics and engagement structures unlike traditional Euclidean clustering that focuses solely on numerical scale.

3.3 Top Influencer Patterns

The top influencers in each cluster were identified based on total engagement to provide deeper insight into performance variation within the segmentation structure. Table 2, Table 3, and Table 4 summarize the highest-performing accounts in each cluster. The results reveal a clear behavioral contrast between the clusters: Cluster 0 (Micro–Mid) is dominated by influencers with stable and consistent engagement, Cluster 1 (Mega) is represented by a single extremely dominant account with exceptionally high reach, and Cluster 2 (Macro) shows influencers with wide exposure but fluctuating engagement intensity.

Table 2. Top 5 Influencers – Cluster 0 (Micro–Mid)

Rank	Influencer	Total Engagement
1	Itsdrryan	167,560,600
2	skincare9711	128,790,100
3	Joshuaomonis	122,415,100
4	Garbashhh	100,392,100
5	Noeminikita	99,654,700

Source: Processed Data (2025)

Table 3. Top Influencer – Cluster 1 (Mega)

Rank	Influencer	Total Engagement
1	l.thomas2020	148,891,100

Source: Processed Data (2025)

Table 4. Top Influencer – Cluster 2 (Macro)

Rank	Influencer	Total Engagement
1	itstherealkimshady	78,345,100

Source: Processed Data (2025)

These results demonstrate that influence within the TikTok endorsement ecosystem is distributed rather than centralized, with each influencer tier performing a distinct strategic role in the marketing funnel. Micro–Mid influencers maintain long-term interaction consistency through authenticity, relational closeness, and conversational content that builds community trust and brand intimacy, making them effective for conversion and retention campaigns where credibility and emotional engagement are vital [23]. Macro influencers, positioned in the middle tier, exhibit flexible engagement across diverse content styles, serving as persuasive amplifiers that balance broad reach with adaptability ideal for mid-funnel goals such as audience persuasion, brand recall, and message amplification. In contrast, Mega influencers function as attention accelerators, utilizing virality and algorithmic amplification to achieve massive visibility and top-funnel awareness, shaping public perception through large-scale exposure [24]. Collectively, these behavioral differences emphasize that influencer selection should be segmented, data-driven, and objective-based, not generalized under a one-size-fits-all approach. A segmentation-based strategy enables efficient resource allocation—deploying Micro–Mid influencers for retention, Macro influencers for persuasion, and Mega influencers for awareness—ensuring higher campaign precision, optimized budgeting, and improved ROI grounded in behavioral evidence rather than superficial popularity metrics.

3.4 PCA Visualization and Interpretation

The PCA (Principal Component Analysis) scatter plot provides a concise two-dimensional visualization of the multivariate engagement space, serving as a key diagnostic tool for validating the cluster structure identified through Spectral Clustering. By projecting standardized feature vectors onto the first two principal components, dimensional complexity is reduced while preserving the main variance patterns, enabling clear observation of between-cluster separation and within-cluster cohesion [25]. Visually, Cluster 0 appears densely concentrated, indicating high homogeneity among Micro–Mid influencers who exhibit moderate play counts, consistent comment rates, and stable, community-driven engagement. In contrast, Cluster 2 occupies a broader, more dispersed area, revealing heterogeneity among Macro influencers, who despite higher visibility differ in the dominant engagement metrics driving their performance, such as shares, comments, or likes, making them functionally diverse yet less homogeneous than Micro–Mid influencers [26]. Meanwhile, Cluster 1 appears visually isolated and positioned far from the others in the PCA space, representing Mega influencers whose engagement magnitudes are significantly higher than average. This separation is not merely due to numerical outliers but reflects fundamental differences in interaction structures, such as disproportionately high shares and likes, highlighting the distinct and intensive engagement behaviors that distinguish Mega influencers from the broader influencer population.

From a methodological standpoint, the Principal Component Analysis (PCA) plot performs two key validation functions for this study [27]. First, it reinforces the validity of the Spectral Clustering results, as the clear spatial separation between clusters in the reduced dimensional space supports the high Silhouette Score and confirms that the clusters arise from meaningful structural differences rather than computational artifacts or overfitting. Second, PCA enhances interpretability by showing which original engagement variables contribute most to cluster differentiation [28]. For example, the first principal component likely represents overall interaction volume encompassing play counts, shares, comments, and likes while the second captures the balance among engagement types, distinguishing groups based on interaction composition rather than scale. To further support these interpretations, presenting component loadings is recommended, as they indicate the relative contribution of each variable and provide transparency in identifying the sources of separation [29]. The PCA pattern also reveals a marketing hierarchy: Mega influencers (Cluster 1) act as top-funnel amplifiers with viral reach, Macro influencers (Cluster 2) provide mid-funnel engagement through diverse mechanisms like shares or sustained attention, and Micro–Mid influencers (Cluster 0) serve as bottom-funnel converters through consistent, high-quality engagement. This structure not only validates the segmentation but also informs multi-tiered campaign strategies. However, because PCA is a linear projection and Spectral Clustering is graph-based and sensitive to affinity scaling [30], further robustness checks such as reporting explained variance and loadings for PC1–PC2, using PC3–PC1 plots, t-SNE/UMAP visualizations, and analyzing winsorized or log-transformed data are necessary to confirm that the observed separations are not solely driven by scale differences [31].

Figure 2. PCA projection of influencers by cluster (k = 3). Two-dimensional principal component projection of standardized engagement features. Points are colored by Spectral Clustering assignments (Cluster 0 = Micro–Mid; Cluster 2 = Macro; Cluster 1 = Mega). The first two principal components capture the dominant variance directions in the data and visually confirm the separation among clusters: Cluster 0 is dense and homogeneous, Cluster 2 is dispersed indicating heterogeneity, and Cluster 1 is isolated due to extreme engagement magnitudes.

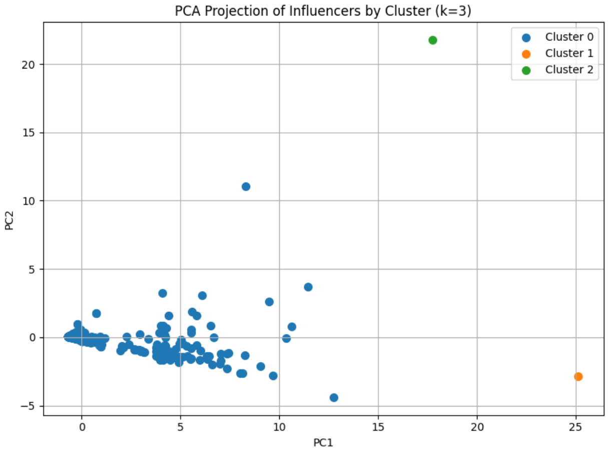


Figure 2. PCA Projection of Influencers by Cluster (k=3)

The visualization confirms the structural validity of the Spectral Clustering results, as the clear spatial separation aligns with the high Silhouette Score, indicating that clusters are behaviorally distinct rather than computational artifacts. The first principal component reflects overall interaction volume, while the second distinguishes influencers based on engagement composition (e.g., shares vs. likes). However, to maintain analytical transparency, several considerations are necessary: PCA, being a linear projection, may not fully capture nonlinear relationships thus, additional visualizations like t-SNE or UMAP are suggested. Reporting the explained variance and loading values for PC1 and PC2 would clarify which features drive separation, while log-transformation or winsorization is recommended to reduce the dominant influence of Mega influencers and ensure that the observed cluster separation is not merely scale-driven.

Although the Silhouette Score (0.9473) indicates strong clustering performance, such a high value warrants careful interpretation. Elevated scores in high-dimensional or small datasets can arise from data redundancy, where similar metrics (e.g., likes and diggs) amplify similarity artificially, or from model overfitting, where the affinity matrix captures noise rather than genuine behavioral variance. To mitigate these risks, this study incorporated dimensionality reduction (PCA), feature normalization, and local similarity constraints to reduce correlation bias. Nevertheless, future research should validate robustness through cross-validation, bootstrapped similarity matrices, or out-of-sample testing to confirm generalizability beyond this dataset.

3.5 Discussion and Interpretation

The discussion emphasizes that influencer effectiveness on TikTok is shaped not merely by follower count but by engagement dynamics, relational intensity, and interaction quality. The three-cluster segmentation Micro–Mid, Macro, and Mega influencers reveals a tiered behavioral structure aligning with the digital marketing funnel: Mega influencers (Cluster 1) drive awareness at the top through viral visibility, Macro influencers (Cluster 2) sustain interest and desire via persuasive and shareable content, and Micro–Mid influencers (Cluster 0) foster conversions through authenticity and community trust. This structure demonstrates that Spectral Clustering offers both computational precision and strategic insight, capturing nonlinear engagement hierarchies and surpassing superficial metrics like follower count. Consistent with prior studies in data-driven marketing analytics, the model enables marketers to predict engagement outcomes, allocate budgets effectively, and align influencer roles with campaign objectives. Thus, a tiered influencer strategy Mega for awareness, Macro for persuasion, and Micro–Mid for conversion ensures balanced resource allocation between visibility and credibility, reinforcing that sustainable digital influence relies on authenticity, emotional connection, and data-driven segmentation.

4. CONCLUSION

This study successfully applied the Spectral Clustering algorithm to segment TikTok influencers into three behavioral groups Micro–Mid, Macro, and Mega influencers based on engagement and interaction metrics. Achieving a high Silhouette Score of 0.9473, the model proved strong cluster cohesion and separation. The key contribution lies in using graph Laplacian eigenvalue decomposition to reveal nonlinear community structures, surpassing traditional linear methods like K-Means or Hierarchical Clustering. The findings show that influencer effectiveness depends more on interaction dynamics, engagement intensity, and relational authenticity than follower count, forming a clear behavioral hierarchy: Micro–Mid for trust and conversions, Macro for reach and persuasion, and Mega for viral awareness. The framework bridges communication theory and computational mathematics, aligning with the AIDA model and offering marketers a data-driven decision-support system for influencer selection, targeting, and budgeting. Future research could extend this model through time-series, sentiment, or cross-platform analysis and integration with supervised learning algorithms to improve predictive performance, reinforcing the study's role in advancing an analytical, transparent, and effective digital marketing ecosystem.

5. REFERENCES

[1] DataReportal, “We Are Social & Meltwater,” 2024. [Online]. Available: https://datareportal.com/reports/digital-2024-indonesia. [Accessed: 15-Aug-2025].

[2] Katadata Insight Center, “Katadata.co.id,” Dec. 2023. [Online]. Available: https://databoks.katadata.co.id/. [Accessed: 15-Aug-2025].

[3] FastMoss, “FastMoss.com,” 2025. [Online]. Available: https://fastmoss.com/. [Accessed: 7-Sep-2025].

[4] A. Rahman, “Data collection, wrangling, and pre-processing for AI assurance,” in *Data Collection and Pre-Processing for AI Assurance*, 2023, pp. 321–338, https://doi.org/10.1016/B978-0-32-391919-7.00022-6.

[5] A. Yusuf and H. Tjandrasa, “Prediksi nilai dengan metode Spectral Clustering dan Clusterwise Regression,” *Jurnal SimanteC*, vol. 4, pp. 1–8, 2020.

[6] T. Zhao and H. C. Y. Zhao, “DSEAGC: Dual-spectral embedding for attributed graph clustering,” *Internet of Things*, vol. 332, p. 101651, 2025, https://doi.org/10.1016/j.iot.2025.101651.

[7] A. Li *et al.*, “Deep spectral clustering network for incomplete multi-view clustering,” *Engineering Applications of Artificial Intelligence*, vol. 148, p. 110387, 2025, https://doi.org/10.1016/j.engappai.2025.110387.

[8] R. Ananda, M. I. A. Latif, A. L. H., and P. S. R. R., “Analisis harga rumah di daerah Jakarta Selatan dengan menggunakan algoritma Spectral Clustering,” *Jurnal Ilmu Komputer Revolusioner*, vol. 8, pp. 65–72, Oct. 2024.

[9] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, pp. 849–856, 2022.

[10] S. Wulandari, “Clustering Microarray Adenoma menggunakan Spectral Clustering dengan algoritma Partitioning Around Medoid (PAM),” *Prosiding Seminar Nasional Sains*, vol. 4, pp. 345–351, 2020.

[11] A. Bose and M. Das, “Enhancing model explainability through S-GuISE: A spectral clustering-guided input sampling scheme for explanation,” *Neurocomputing*, vol. 650, p. 130750, 2025, https://doi.org/10.1016/j.neucom.2025.130750.

[12] J. Huang *et al.*, “A hybrid framework for assessing regional inertia estimation in bulk power systems using COI-driven spectral clustering,” *International Journal of Electrical Power and Energy Systems*, vol. 169, p. 110457, 2025, https://doi.org/10.1016/j.ijepes.2025.110457.

[13] X. Chen *et al.*, “A novel multi-means joint learning framework based on fuzzy clustering and self-constrained spectral clustering for superpixel image segmentation,” *Computers and Electrical Engineering*, vol. 125, p. 109872, 2025, https://doi.org/10.1016/j.compeleceng.2025.109872.

[14] J. Zhang *et al.*, “ClusterRiceNet: A novel rice seed variety classification network based on hyperspectral imaging and spectral band clustering,” *Knowledge-Based Systems*, vol. 305, p. 112890, 2025, https://doi.org/10.1016/j.knosys.2025.112890.

[15] D. B. Lodianto, “Spectral Graph Theory for Community Detection in Social Networks,” *Institut Teknologi Bandung*, 2024.

[16] J. Wang and A. Robles-Kelly, “Spectral contrastive clustering,” *Pattern Recognition*, vol. 166, p. 111671, 2025, https://doi.org/10.1016/j.patcog.2025.111671.

[17] Z. Zhang, J. Xu, W. Zhao, J. Xie, and F. N. Xiang, “Spectral ensemble clustering from graph reconstruction with auto-weighted cluster,” *Pattern Recognition Letters*, vol. 196, pp. 243–294, 2025, https://doi.org/10.1016/j.patrec.2025.03.015.

[18] L. Yang, X. Wang, and G. W. F. Alshami, “SCDFL: A spectral clustering-based framework for accelerating convergence in decentralized federated learning,” *Computer Networks*, vol. 271, p. 111615, 2025, https://doi.org/10.1016/j.comnet.2025.111615.

[19] X. Zhao and Q. C. B. Yan, “Individual fair fuzzy C-means clustering via density-adaptive spectral regularization,” *Neurocomputing*, vol. 651, p. 130794, 2025, https://doi.org/10.1016/j.neucom.2025.130794.

[20] S. Zhang, C.-G. Li, X. Qian, R. Xiao, and J. G. Wang, “Neural normalized cut: A differential and generalizable approach for spectral clustering,” *Pattern Recognition*, vol. 164, p. 111545, 2025, https://doi.org/10.1016/j.patcog.2025.111545.

[21] R. Feng *et al.*, “Hypergraph dismantling with spectral clustering,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 150, p. 108975, 2025, https://doi.org/10.1016/j.cnsns.2025.108975.

[22] A. G. J. T., S. T.-S. H., and A. M. A. Siddique, “Big data analytics in food industry: A state-of-the-art literature review,” *npj Science of Food*, vol. 9, 2025, https://doi.org/10.1038/s41538-025-00394-y.

[23] U. von Luxburg, “A Tutorial on Spectral Clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2020.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an Algorithm,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, pp. 849–856, 2020.

[25] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2020.

[26] A. Kumar and R. Kannan, “Clustering with Spectral Norm and the k-Means Algorithm,” *Foundations of Computer Science (FOCS)*, pp. 299–308, 2020.

[27] A. Avrachenkov, M. Bobu, and J. Dreveton, “Higher-Order Spectral Clustering for Geometric Graphs,” *Journal of Fourier Analysis and Applications*, vol. 27, no. 6, pp. 1–27, 2021, doi: 10.1007/s00041-021-09825-2.

[28] A. Dall’Amico, R. Couillet, and N. Tremblay, “A Unified Framework for Spectral Clustering in Sparse Graphs,” *Journal of Machine Learning Research (JMLR)*, vol. 22, no. 261, pp. 1–55, 2021.

[29] P. Macgregor and H. Sun, “A Tighter Analysis of Spectral Clustering, and Beyond,” in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, Baltimore, MD, USA, 2022, pp. 15083–15106.

[30] A. Cerro, S. Dasmahapatra, A. Day-Hall, L. Moretti, *et al.*, “Spectral Clustering for Jet Physics,” *Journal of High Energy Physics (JHEP)*, vol. 2022, no. 2, pp. 165–187, 2022, doi: 10.1007/JHEP02(2022)165.

[31] H. Chen, H. Ye, and C. Li, “Spectral Clustering Community Detection Algorithm Based on Point-Wise Mutual Information Graph Kernel,” *Entropy*, vol. 25, no. 12, pp. 1617–1634, 2023, doi: 10.3390/e25121617.