Unsupervised YouTube Video Segmentation of "Bendera One Piece" Content Using Medoid-Based Clustering with Statistical Significance Testing

W Budiaji[1,*], P P Kumenap[1], M F R Delano[1], F B Wijaya[1], R A Riyanto[2,3]

[1] Statistics Department, University of Sultan Ageng Tirtayasa, Cilegon 42435, Indonesia
[2] Food Technology Department, University of Sultan Ageng Tirtayasa, Cilegon 42435, Indonesia
[3] Department of Biotechnology, Osaka University, Osaka 565-0871, Japan

*Corresponding author's email: budiaji@untirta.ac.id

Abstract. The curse of dimensionality and sparsity are well-documented phenomena in applied statistics where the data's dimensionality (number of features) far outnumbers the observations. This work aims to present an integrated applied statistics framework to distill semantic structure from high-dimensional data by combining pre-processing, dimensionality reduction via principal component analysis, medoid-based clustering (partitioning around medoids and simple k-medoids), and a modified Statistical Significance Clustering (SigClust) test for validation and inference in the context of viral media. In this case study, we demonstrate an approach that segments and interprets YouTube videos from the lens of the Indonesian viral media "Bendera One Piece" through its user commentary. The PCA-based dimensionality reduction helped resolve the curse of dimensionality, where the first principal component alone explained 80% of the variance in text-based features and captured a dominant socio-political pattern. Internal validation and the SigClust test agreed on the presence of a statistically significant three-cluster solution that could be labelled as the audiences of "Pop-Culture Enthusiasts", "Cautious Observers", and "Political Protesters". The study presents the utility of integrating established statistical methods with a modified validation step for high-dimensional text data analysis and pattern recognition.

Keywords: Applied Statistics, High-Dimensional Data, Medoid-Based Clustering, Principal Component Analysis, Statistical Significance Testing.

1. Introduction

YouTube has been established as a de-facto go-to platform for visual media today. It is both an engine of global entertainment and an important hub for spreading information and ideas [1]. It allows people to upload a range of content, from educational resources [2], [3] and news [4] to culture [5] and populism [6]. One of the unique characteristics that differentiates the platform from conventional broadcast media is its inherent interactivity, most notably via the video comments section. In this way, it functions as a direct feedback loop and a virtual forum, whereby viewers are empowered to participate and interact with content beyond simple passivity [6]. Viewers post their reactions, emotional responses,

comments, opinions, and even feedback for other viewers. By its nature, this is a large archive of user-generated text that is reflective of the viewers' understanding of the content, the positive/negative tone of their responses, and the social meaning ascribed to a particular video.As a result, the platform has generated a high-dimensional information landscape at a global scale.

A case in point is the video and image macro of Indonesians flying 'One Piece' flags on their vehicles in the run-up to Indonesia's Independence Day celebrations [7]. The popularity of the practice has reached viral trend status. Flying Bendera One Piece has become a way for Indonesians to express their feelings of discontent about the state of the nation and protest inequality. It is also more broadly understood as a symbol of protest and social aspiration [8]. The videos uploaded to YouTube from the related Bendera One Piece trending topic and others such as the actions have thousands of comments and can be assumed to be representative of the Indonesian audience's mood and thinking. In other words, we have access to a high-dimensional, Indonesia-wide public opinion of a globally popular cultural phenomenon and entertainment.

The task is, however, not straightforward. On the one hand, we have a real-world manifestation of a pop culture product that has become well-established in the country. For instance, One Piece is one of the most firmly planted shonen anime in the Indonesian collective consciousness [8], [9]. In that sense, there are several independent instances of the topic which need to be identified and separated. On the other hand, it has also been used as a channel of nationalistic pride and a general protest against social inequities, which is not a One Piece show by itself. We aim to segment the YouTube videos into semantically similar sets of videos that reflect what are effectively audience segments. The issue is how to achieve this in an unsupervised manner.

A proposed approach is to use the video comments to perform segmentation. The data associated with the problem is obviously high-dimensional. The number of observations (n) is the number of unique videos. The number of features (p) is the number of unique word tokens in its comments. Thus, in all practical settings, $n << p$. This makes the problem of curse of dimensionality [10]. The dataset is so high-dimensional that any standard measure of distances (e.g., Euclidean distance) between observations will be lose their meaning, because points in such spaces are too far apart to compare. For clustering, this means that any standard clustering technique such as k-means [11] is susceptible to delivering random results.

To address this research problem, we implemented a tailored statistical pipeline that integrates several established techniques. There are four steps involved. 1) Structuring the raw, unstructured data from video comments using text preprocessing and tokenisation [12]. 2) Dimensionality reduction is required to reduce the thousands of unique word tokens to a low-dimensional, latent feature space (vector subspace) that preserves as much variation as possible, this can be achieved by using Principal Component Analysis (PCA). 3) Cluster analysis or pattern recognition then needs to be used to implement a medoid-based algorithm that will segment the videos on the set of PCs based on their similarity. 4) The resultant clusters need to be validated statistically to confirm that they exist and are not random noise using statistical significance testing (SigClust). A key adaptation in this pipeline is modifying the standard SigClust test. It now accepts cluster labels from medoid-based algorithms and supports more than two clusters, which improves its usefulness.

The approach in this article was informed by and contributed to a number of established statistical areas. The first step of the text pre-processing involved case folding, tokenisation, stopword removal, and filtering [13], [14]. This is considered textbook and foundational in the field of text mining or natural language processing (NLP). The first substep is required because computational tools, like R, are case-sensitive to upper/ lower letters [15]. Moreover, the whole purpose of this step and its further steps is to reduce noise in the data. Thus, tokenisation is critical, as it directly affects the entire further analysis [16]. Stopword removal and stop-filtering, meanwhile, are common NLP pre-processing steps to improve model fit [17], [18].

The need to use dimensionality reduction is a recognised challenge in applied statistics and machine learning. The statistical challenge of high-dimensional data is well-established, with the resultant statistical data being sparse, and hence the poor performance of many standard statistical and machine learning algorithms [11]. Principal Component Analysis, meanwhile, is standard linear algebra, implemented in R as an unsupervised factor extraction procedure via matrix factorisation. It will project a set of correlated variables onto a new set of uncorrelated variables (Principal Components or PCs) ranked by the amount of variation captured [19]. It is, by definition, an algorithm for dimensionality reduction and represents the linear projection of the global variance in the original space onto a lower dimension.

For clustering, we adopt the k-medoids algorithm, a robust alternative to k-means. While k-means is among the top 10 data mining algorithms, it is particularly suited only to working with Euclidean distances and is sensitive to outliers. A k-medoids algorithm, on the other hand, is a generalised version of k-means in which the cluster centres are actual data points (medoids) and a general distance measure can be applied [20]. Cluster validation is then part of the standard toolkit for unsupervised learning. The silhouette and medoid-based shadow value indices are both standard internal cluster validation indices measuring cluster cohesion and separation [21]. However, internal indices do not provide p-values for significance. This is what the SigClust is designed to do, to provide formal statistical testing of Statistical hypothesis for the clustering result [22], [23].

While SigClust is designed for statistical testing, it is also linked to k-means and hierarchical clustering. Our work addresses a gap by changing the SigClust procedure to accept external cluster labels from medoid-based clustering. We also implement a post-hoc correction for multi-cluster solutions. This approach offers a more thorough and adaptable validation method.

In the "Bendera One Piece" phenomenon, which is a significant cultural event in Indonesia, no previous study has used an unsupervised statistical method to segment and measure the various viewers behind this viral trend. Our research addresses this gap by offering the first data-driven, statistically confirmed audience segmentation for this event.

Following the research problem and the identified gap in integrated, statistically validated text segmentation pipelines, this study has two main objectives:

1. To develop and validate a strong statistical pipeline for high-dimensional text segmentation, we will integrate Principal Component Analysis (PCA) with medoid-based clustering. We will also modify the Statistical Significance Clustering (SigClust) test for thorough validation.
2. This pipeline will be used to segment and understand the audience of the "Bendera One Piece" YouTube phenomenon. Through this, we aim to provide an empirical analysis of the various socio-political and cultural narratives within this viral discussion.

2. Research Method

2.1. Data acquisition and selection criteria

The data for the empirical section of this paper were harvested through an automated process from the YouTube application as shown in the Results. The scraping of the data was done in compliance with the service's terms of use by using the official public YouTube Data API v3 from Google via the R tuber package [24]. It was performed on the dates of 16 August and 17 August 2025 before 10: 00 AM (every 17 August is the holy hour of nation's independence day). To ensure the videos were genuinely part of the Indonesian discourse, a search query with the keyword string "bendera one piece" was applied. "bendera" is the Indonesian word for flag. The search query was specific to primary video data (`type = "video"`) sorted by view counts in descending order (`order = "viewCount"`). From the resulting list, the N = 100 videos with the most views were selected and then filtered to include only those with at least 100,000 views (`viewCount > 100,000`). This threshold was chosen to identify content that reached a large audience and could be seen as "viral" or having a big impact on the platform.

As a result, the final corpus contained a total of N = 34 videos that met all criteria with their associated comments, which would serve as the raw text input for our analysis.

2.2. Text data pre-processing: from raw text to structured tokens

The unfiltered text (raw strings of characters) from comments on videos is an unstructured corpus. As such, it is in a form not yet suitable for analysis and must be transformed by a series of pre-processing operations into a structured and quantifiable set of semantic tokens. This step is a critical component of the entire pipeline that has a significant impact on the downstream noise-to-signal ratio and accuracy of the clustering.

2.2.1. *Case folding and non-textual character removal*: This is the first pre-processing stage and standardizes the case by converting all alphabetical characters to lowercase (equation 1). This is a necessary normalization step as by default most programs read and treat tokens with case-different alphabets as distinct tokens. For example, in the original unfiltered text, the tokens "Bendera", "bendera", and "BENDERA" would each be considered separate features/variables with their respective term frequency values. This multiplies the dimensionality of our term space and distorts the true term frequency measures of each. This problem can be avoided if we enforce a standardized case using:

$$S_{folded} = lower\ (S) \quad (1)$$

where S denotes the given input text as a string, lower denotes a string-processing function that outputs the case-folded version of the text. Simultaneously, all numbers, punctuation, and special characters were removed, which generally hold little to no meaningful information for our topic-based clustering and is considered noise.

2.2.2. *Tokenization*: The pre-processed input text string, S folded, is then passed through a tokenization algorithm. The purpose of this function is to subdivide the standardized text string (now without structure) into individual units of semantic meaning called tokens which are most often individual words (equation 2). The algorithmic process of Tokenization is formalized as the following:

$$T(S_{folded}) = \{t_1, t_2, t_3, \dots, t_m\} \quad (2)$$

where each t with a subscript i represents the i-th token in the resultant list/set of m tokens produced by applying function T to the text string. This results in a bag-of-words (BOW) representation of the previously unstructured and untokenized string S. The output of T represents the deconstructed lexical elements from the unstructured input that serves as the raw input for our downstream statistical modelling.

2.2.3. *Stopword removal*: The next and penultimate pre-processing step to apply to the tokens is Stopword Removal. Stopwords are generally the most high-frequency function words in a language that often lack substantive semantic differentiation within a body of text (e.g., "yang", "dan", "di" in Indonesian) and hence, little to no value for a clustering method concerned with identifying topical distinctions. They are usually not useful content-bearing keywords and can be effectively removed (equation 3). For our application, a general predefined stopword list, Lstop, provided by the `stopwords` R package [25] for Bahasa Indonesia was used. This is implemented as the following formal operation:

$$T_{filtered} = \{t_1 |\ t_1 \epsilon\ T \wedge t_1 \notin\ L_{stop}\} \quad (3)$$

This step can be considered a form of feature selection that helps to dramatically reduce dimensionality and focus the subsequent feature set on more content-bearing keywords and terms.

2.2.4. *Lexical filtering*: The final data pre-processing stage, Filtering, is a rudimentary lexical filter based on the length of tokens to remove possible noise and residuals from previous steps. Any tokens with a character length of less than 3 were removed as they are often typos or miswritten words, informal abbreviations, or non-content-bearing interjections like "oh" "ah". Those above

15 were also removed, as they often represented URLs, severe misspellings, or non-lexical strings (equation 4). The formula of this filter function, F, is:

$$T_{final} = \{t_1 | \, t_1 \in T_{filtered} \wedge 3 \leq len\,(t_1) \leq 15\} \quad (4)$$

The resultant Tfinal, the filtered token list, for all comments from all N = 34 videos is the desired final lexical universe for this study.

2.3. *Feature matrix construction and dimensionality reduction*

.

2.3.1. *Document-Term Matrix construction*: The pre-processed tokens were then used to generate a (sparse) Document-Term Matrix (DTM), **X**. In the resultant matrix, the 34 rows are the video files (documents) and the p = 3,618 initial columns are the individual tokens (terms). Each entry in the matrix, xij, counts the frequency of term j in video i (document). By construction, this is an extremely sparse matrix with very high dimensionality, where the number of documents n = 34 << p = 3,618. This would make any direct clustering impossible. Dimensionality reduction is required first.

2.3.2. *Feature selection and semantic aggregation*: To move from our starting DTM towards a relevant feature set that best characterizes this corpus, a feature selection process must be performed. First, a proportional threshold was selected and applied. Any feature (word/token) that did not appear at least in 50% of all documents (videos) was filtered out. Thus, a word would have to appear in at least 0.5 times 34 = 17 videos to be retained. This ensures that we are only including terms that are common to the general population of videos rather than some idiosyncratic outlier. This step was used to reduce the feature space to 142 terms.

A secondary manual lexicon refinement was conducted to remove all words that made it past the frequency threshold but were semantically no meaning (e.g., “gini”, “gitu”). Finally, we aggregated frequencies of the synonymous terms that were clearly identified as thematically similar/identical (e.g., all frequencies of both “bilang” and “ngomong” are added to a single meta-feature that semantically represents the concept of “speaking”). This second aggregation step then completed the feature selection process, leaving us with a powerful set of (p = 109) content-bearing keywords which produced the final DTM **X** of size 34 times 109.

2.3.3. *Principal Component Analysis (PCA)*: Principal Component Analysis (PCA) was performed to project the data to a subspace. It is an orthogonal linear transformation that projects the correlated features onto a new uncorrelated feature subspace with dimensions equal to the number of original features [26]. These new uncorrelated features are the so-called Principal Components (PCs) and are ordered such that the first PC accounts for the highest possible variance, the second the next, and so forth.

Formally, the first principal component z1 is the linear combination of the original features (equation 5). It maximizes the sample variance:

$$\boldsymbol{z}_1 = \boldsymbol{X}^* v_1, \text{ where } v_1 = \arg\max_{\|v\|=1} Var\,(\boldsymbol{X}^* v) \quad (5)$$

v1 is the eigenvector of the covariance matrix $\boldsymbol{C} = \frac{1}{n-1}(\boldsymbol{X}^*)^T \boldsymbol{X}^*$ corresponding to the largest eigenvalue λ1 and the proportion of the total variance explained by the k-th PC is given by the formula $\lambda_k \left(\sum_{j=1}^{p} \lambda_j\right)^{-1}$.

The number of PCs, k, was chosen using a scree plot (graph of eigenvalues λk vs k). This results in projecting the original data onto a lower-dimensional subspace Z of PC scores of dimension 34 times k, while preserving as much of the global variance structure as possible. It is now orthogonal and free of the multicollinearity and noise that plagued the original feature space. PCA was chosen for dimensionality reduction instead of other modern techniques like word embeddings because it works directly on the DTM. This enables us to identify the elements that PC1 captured. PCA is specifically designed to find the directions of maximum variance in a

dataset. This fits well with the goal of segmenting videos based on the most dominant themes in the corpus.

2.4. *Clustering Methodology and Validation*

Following dimensionality reduction, we employed medoid-based clustering to segment the videos. This approach was selected over topic modelling methods like Latent Dirichlet Allocation (LDA) because the primary objective was to partition the videos themselves into discrete, semantically coherent viewer segments, rather than to discover mixed-membership topics within the corpus. Clustering the principal component scores allows for a clear, hard assignment of each video to a single segment, which is more directly interpretable.

2.4.1. *Medoid-Based Clustering Algorithms*: The resulting matrix Z of PC scores was clustered/partitioned using a medoid-based algorithm. This method is distinct from the standard k-means type algorithm, which uses artificial means as the center of clusters. A medoid-based algorithm, on the other hand, selects actual data points from the dataset as the medoids of the clusters. The choice of a medoid-based clustering algorithm was driven by the specific needs of this study. While k-means is a popular method, it has two main limitations for this application. First, it uses the mean of observations as the cluster center, which is very sensitive to outliers. Second, it is primarily designed to work with Euclidean distance. In contrast, medoid-based algorithms have two main advantages. They are robust against outliers and noise, and they can use any valid distance measure [20]. This allows for an empirical choice of the best metric to use. For a given number of k clusters, the problem of medoid-based algorithm can be formalized as: Goal: Find the set of $M = \{m_1, m_2, m_3, \dots, m_k\}$ of data points that minimize the total within-cluster dissimilarity (equation 6). The total within-cluster dissimilarity is defined as:

$$W = \sum_{i=1}^{k} \sum_{z \in C_i} d(z, m_i) \quad (6)$$

where Ci is the set of all points in cluster i, m is the medoid of cluster i and d(.) is a distance function. In this study, two separate algorithms were used for verification:

1. Partitioning Around Medoids (PAM)[27]: A more comprehensive algorithm that exhaustively minimizes the sum of dissimilarities between points and their closest medoid.
2. Simple K-Medoids (SKM)[28]: A k-means-like, but instead of using means for the cluster center uses a medoid-based approach.

The cluster performance was evaluated across three different metrics of distance to identify the sensitivity of the final output to this critical input parameter. The distance functions used were Euclidean (equation 7), Squared Euclidean (equation 8), and Manhattan (equation 9). The general formula for a pair of points, a and b, in any k-dimensional space in Z is as follows:

Euclidean: $$d(a, b) = \sqrt{\sum_{j=1}^{k} (a_j - b_j)^2} \quad (7)$$

Squared Euclidean: $$d(a, b) = \sum_{j=1}^{k} (a_j - b_j)^2 \quad (8)$$

Manhattan: $$d(a, b) = \sum_{j=1}^{k} |a_j - b_j| \quad (9)$$

2.4.2. *Internal cluster validation*: The internal validation of the final output was conducted to verify and confirm the appropriate number of clusters (k) in a range from 2 to 10 and the quality of the partition. For this purpose, silhouette (equation 10) and medoid-based shadow value (equation 11) metrics [21] were used. These are two specifically designed indices to measure the separation between medoids. They are each defined as:

Silhouette coefficient:

$$si\ (x) = \frac{(b_x - a_x)}{\max(a_x, b_x)} \quad (10)$$

where ax and bx are the average distances of an object x to all the other objects in the same cluster and to all objects in the closest other cluster. The global average silhouette width was used.

Medoid-Based Shadow Value:

$$msv\ (x) = \frac{d\left(x,m'(x)\right)-d(x,m(x))}{d(x,m\prime(x))} \quad (11)$$

where d(x, m(x)) is the distance between object x to the first closest medoid and d(x, m`(x)) is the distance between object x to the second closest medoid. The best performing clustering configuration for each algorithm, distance, and k is selected as the final one (in terms of both silhouette or shadow value).

2.4.3. *Statistical significance testing via the SigClust*: To further ensure that the clusters that were found are not purely from spurious noise that is known to plague high-dimensional data, we applied a statistical test to reach this conclusion. This test was applied using the `sigclust` package [22]. Formally, this is applied as follows:

The SigClust is applied to the original high-dimensional dataset **X**, with the cluster labels obtained C from the optimal medoid-based partition on the lower-dimensional PCA-reduced data Z. The original data does not have to be normal for SigClust to work. The test's goal is to discover whether the data can be described by a single normal distribution.

The SigClust tests the null hypothesis:

H0: The data is from a single d-dimensional Gaussian distribution (i.e., no clusters)

H1: The data comes from a non-Gaussian distribution (e.g., mixture, many clusters)

It then calculates a p-value based on a cluster-specific index (e.g., 2-means CI) and its distribution under the null hypothesis, simulated using Monte Carlo. If the resulting p-value is significant ($p<0.05$), then it can be stated with a certain confidence level that the clusters are real and not a mere artifact of the high-dimensional noise.

2.5. Modification and application of SigClust for validation

By default, the SigClust algorithm uses 2-means clustering to generate the cluster membership [22], which becomes the input for calculating the cluster index (CI) as well as simulating the null distribution. In this study, the modification applied to it is that the cluster membership to be used for the SigClust test was instead obtained from one of the two medoid-based algorithms (either PAM or SKM) run in an earlier step. The motivation of this was to provide some variation of distances that can be used.

However, as exploratory clustering analysis using this data can result in the formation of more than 2 clusters, the standard SigClust approach of only looking at 2 at a time would be insufficient. The solution to this problem was to expand the scope of the test by running the SigClust on all possible cluster pairs (all possible combinations). For example, if there are three resulting clusters (A, B, and C), then significance tests are run separately between pairs (A vs. B), (A vs. C) and (B vs. C). This approach allows for the validation of the boundary between all clusters. However, performing such multiple tests at once introduces a higher risk of Type I Errors (false positives, i.e., declaring that cluster separation is significant when it is not). In other words, it is more likely to reject the null hypothesis when it should not. To mitigate this effect, the level of significance that is used is corrected using post-hoc nonparametric pairwise multiple comparison tests Holm method [29]. This Family-Wise Error Rate (FWER) correction adjusts the critical p-value that is to be used for each pairwise test based on the number of total tests being performed. It is also less sensitive to the violations of assumptions [30]. With this change to the algorithm, the SigClust provides a strong statistical basis for the validity of the generated segmentation.

3. Result and Discussion

3.1. Dimensionality reduction

To address the curse of dimensionality in the 34x109 matrix, PCA was conducted. The initial scree plot of the eigenvalues indicated that the first two components were able to account for a very large amount of the variation within the dataset, namely 93% of the variance. The score plot of the first two principal

components, however, identified one main outlier (Video number 23) that is located at an extremely extreme position (far from the main cluster of points in the data) (Figure 1). To avoid one video to have a disproportional influence on the dimensionality reduction, the 23th row was dropped from the dataset.

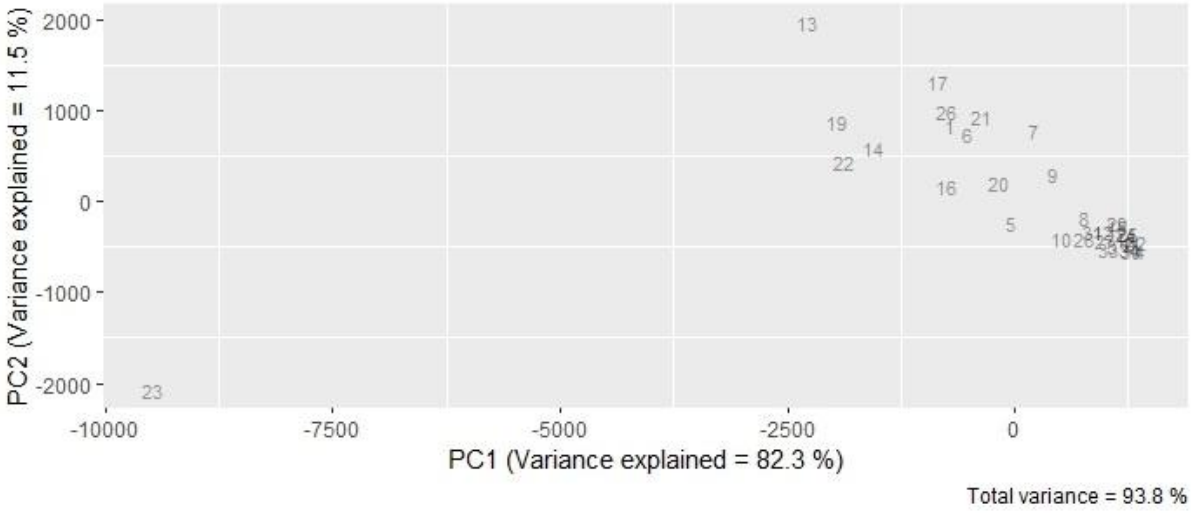

Figure 1. The first two principal component plot.

PCA was then conducted again, this time for the 33x109 matrix. The scree plot was then evaluated again using the elbow method (Figure 2), which now suggested retaining 4 components instead (as the gain in explained variance starts to taper after the fourth component). As was previously the case, the first component (PC1) is clearly dominant, with an eigenvalue close to 80% (Figure 2). This indicates that it is the single strongest, most dominant thematic pattern underpinning the entire dataset of YouTube comments.

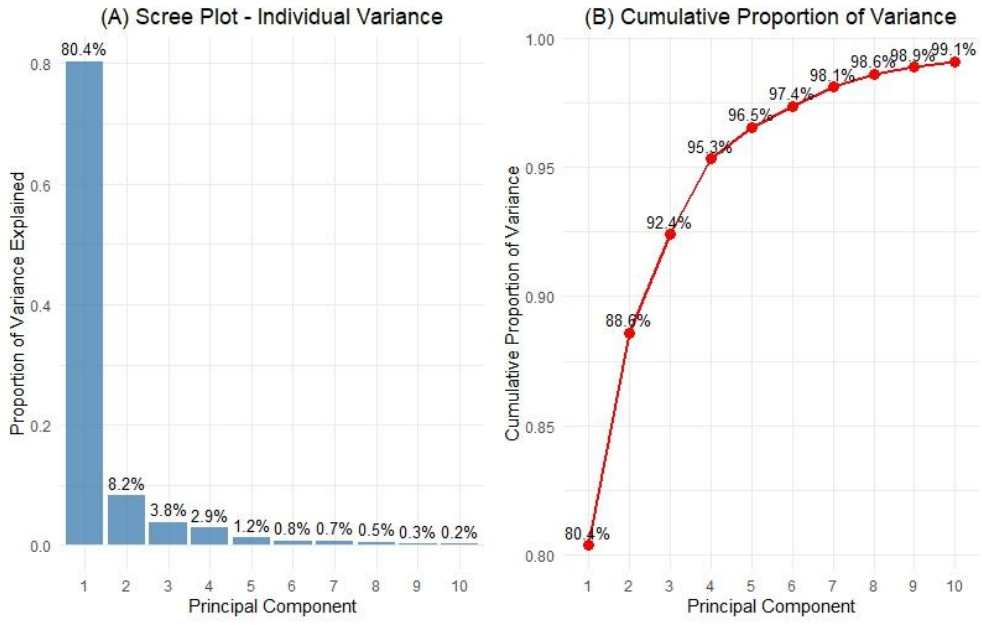

Figure 2. Scree plot (A) and cumulative proportion of variance (B) of the 10 first principal components.

In order to interpret this main thematic pattern, the individual component loadings are taken into account. PC1 is characterized by having extremely high loadings for the words “pemerintah”, “rakyat” and “bendera”. This unique combination is semantically very powerful, and makes it extremely likely that the first main axis of variation of the social narrative (and thus the most salient theme to the audience) has to do with the subject of social commentary. The co-occurrence of the terms in this first component effectively reframes the “Bendera One-Piece” as not only an act of fandom, but rather, in the eyes of the largest share of audience, as a socio-political symbol or statement. It seems very likely that the first main component of variation in audience comments is, in fact, the underlying presence of a protest story or message communicated from “rakyat” to “pemerintah”. In this way, the dimensionality reduction has both successfully boiled down the high-dimensional, messy text data to its most essential, latent structure, and also has made the overall socio-political underpinning of the viral event immediately obvious.

3.2. Optimal cluster via internal validation

Internal validation was conducted to select the most robust and separated cluster structure. The results, visualized in Figure 3, strongly indicated that Squared Euclidean distance should be chosen as the distance metric to be used, as it outperformed the other metrics in terms of internal validation index scores at every number of clusters (k) and with both partitioning algorithms. This superior performance indicates that it was most effective to make larger differences between observations larger, as is the case in the squared distance.

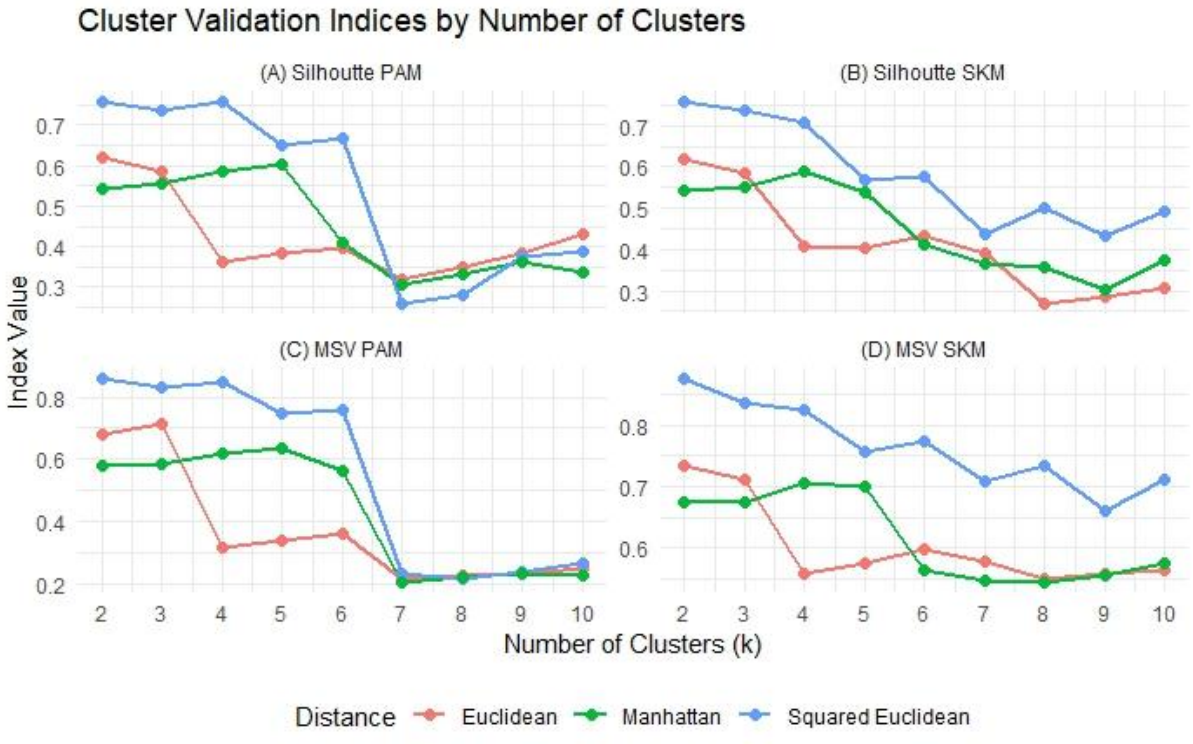

Figure 3. Cluster internal validation indices plot of silhouette in PAM (A) and SKM (B) algorithms and medoid shadow value (MSV) in PAM (C) and SKM (D) algorithms for (k = 2 to k = 10)

In regards to the number of optimal clusters (k), the internal validation indices (Silhouette Width and Medoid-Based Shadow Value) indicated optimal k is 3 or 4 (Figure 3). This k value indicates that partitioning the 33 videos into three/ four distinct clusters should create the best balance between high

similarity within clusters, and low similarity between the different groups. These two cluster solutions were then selected as they provide the most semantically interpretable and well-separated clusters.

As seen in Figure 3, both partitioning algorithms (Partitioning Around Medoids (PAM) and Simple K-Medoids (SKM)) obtained their best scores when used with Squared Euclidean distance, outperforming the same algorithms run on standard Euclidean and Manhattan distances. The two algorithms were directly compared and the performance of the two algorithms is nearly identical (Table 1); however, PAM is consistently rated higher than SKM on both internal validation indices. Due to this performance and its reputation as the popular algorithm in this category, PAM with Squared Euclidean distance was selected as the final and optimal combination for clustering. The selected combination was confirmed to be the most clearly separated, well-partitioned solution in the subsequent statistical analysis.

Table 1. Cluster internal validation indices in PAM and SKM algorithm (k = 2 to k = 5)

Index	Distance	Number of clusters (k)			
		2	3	4	5
Silhouette PAM	Euclidean	0.62	0.58	0.36	0.39
	Squared Euclidean	0.76	**0.74**	**0.76**	0.65
	Manhattan	0.54	0.55	0.58	0.6
MSV PAM	Euclidean	0.68	0.71	0.32	0.34
	Squared Euclidean	0.86	**0.83**	**0.85**	0.75
	Manhattan	0.58	0.59	0.62	0.64
Silhouette SKM	Euclidean	0.62	0.58	0.41	0.4
	Squared Euclidean	0.76	**0.74**	**0.71**	0.57
	Manhattan	0.54	0.55	0.59	0.54
MSV SKM	Euclidean	0.74	0.71	0.56	0.58
	Squared Euclidean	0.88	**0.84**	**0.83**	0.76
	Manhattan	0.68	0.68	0.71	0.7

3.3. Statistical significance of the identified clusters

To move beyond internal metrics, and thus provide statistical evidence for the existence of clusters, a modified SigClust analysis was performed. The generated memberships from the selected combination (PAM with Squared Euclidean distance) were provided as direct input to the significance test, for k=3 as well as k=4. This represents a key methodological modification. The default SigClust function in R package performs the test based on k-means partitioning using Euclidean distance. The adaptation allows for the SigClust to be performed on cluster memberships directly, instead of performing the clustering part itself. This makes it possible to validate cluster membership generated by any combination of partitioning algorithm and distance metric. This is a key improvement, as it means that the specific clustering method that turned out to be optimal for the data does not need to be substituted for the k-means implementation, but can be validated directly instead.

The SigClust test was performed on all possible pairwise comparisons of the three clusters. For each pair of clusters, the null hypothesis (H0) that the two were sampled from a single multivariate Gaussian distribution. The successful rejection of the null hypothesis for all cluster pairs gives statistical support for the pipeline developed for Objective 1. As a correction for the increased chance of Type 1 errors due to multiple testing, the resulting p-values were corrected using the Holm method, a conservative step-down adjustment. The results, displayed in Table 2, are clear. In the k = 4 solution, a pairwise comparison could not be rejected after adjustment, and was not sufficient to provide statistical evidence that the four groups could be considered distinct populations.

Table 2. Statistical significance test for k = 4 and k = 3

	Number of clusters k = 4						Number of clusters k = 3		
	1 vs 2	1 vs 3	1 vs 4	2 vs 3	2 vs 4	3 vs 4	1 vs 2	1 vs 3	2 vs 3
P value	0.009	0.352	0.007	0.005	0.000	0.000	0.006	0.000	0.000
Holm adjusted	0.025	0.05	0.008	0.008	0.008	0.008	0.0167	0.0167	0.0167
Significant	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

In the k = 3 solution, however, all pairwise comparisons (Cluster 1 vs 2, Cluster 1 vs 3, and Cluster 2 vs 3) resulted in statistically significant p-values ($p<0.05$) even after the conservative Holm adjustment. This allows for the confident rejection of the null hypothesis for every cluster pair. This, in turn, provides statistical evidence that the three identified segments do, in fact, form distinct substructures within the data, and are not merely the products of high-dimensional noise or the clustering algorithm itself. The final, statistically validated segmentation thus consisted of 3 clusters, of sizes 8, 21, and 4, videos respectively.

This step is the study's methodological application contribution. By allowing for integration of medoid-based clustering result (PAM with Squared Euclidean distance) into the SigClust and subsequent multiple-testing correction, instead of hierarchical clustering [23], we have, in fact, implemented a more flexible pipeline for validation. This adds more validation to the pipeline prior to the interpretation step and guarantees that the interpretation is not based on algorithmic partitioning alone, but also on a statistical test confirming that the clusters are "really there."

3.4. Interpretation of video segments

Applying the final, statistically validated pipeline (Objective 2), three-cluster solution provides a clear and coherent segmentation of the audience, with each group viewing the "Bendera One-Piece" through a distinct, coherent lens, defined by their own use of distinctive vocabulary.

3.4.1. Cluster 1: The Pop-Culture Enthusiasts (n = 8).

Videos in this segment are overwhelmingly defined by keywords related to entertainment, humor, and the One Piece story itself. Words such as suka (like), anime, kocak (funny), roger (directly referencing the Pirate King Gol D. Roger), berani (brave), and onepiece (Figure 4) are defining vocabulary for the segment, indicating a dominant focus on the source material and, by extension, the global anime narrative. The overall sentiment of the cluster is positive, and the videos can be clearly labeled as "The Pop-Culture Enthusiasts". This group is apolitical in their focus on the event as an act of fandom. They are taking a very popular, globally produced and consumed media, using One Piece-related symbols in a creative way, and bringing them into the local context for fun, entertainment and community.

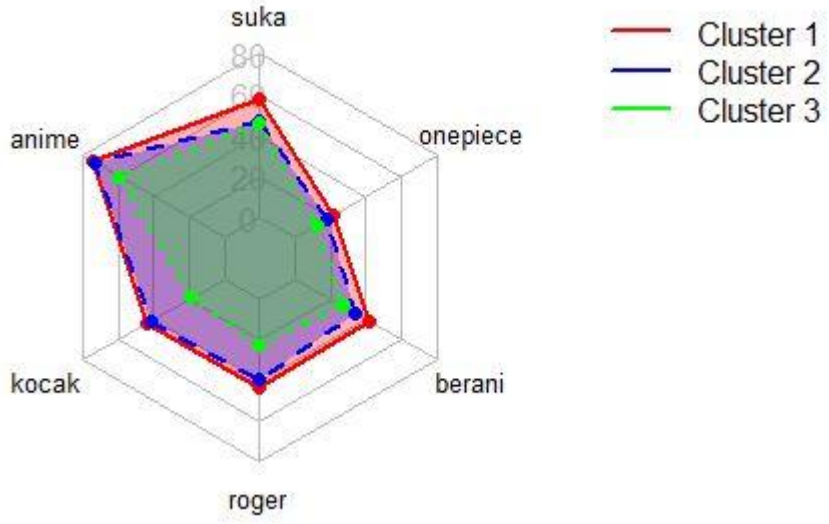

Figure 4. Strong words of Cluster 1 Pop-cluster enthusiasts in radar plot

3.4.2. Cluster 2: The Cautious Observers (n = 21)

As the largest of all segments, the general sentiment of this cluster is driven by a very small set of a few extremely strong, abstract nouns. Words such as partai (political party), sadar (awareness/consciousness), rumah (home/house), and nyata (real) indicate a strong presence of a theme of "awakening to reality" (Figure 5). The naming "The Cautious Observers" accurately encapsulates the general discourse of this large group of audience members. Their comments are largely focused on the "real" world, and the "awareness" of the potential effects of the state of the political "party" on the people's "home". While this group is not as oppositional or direct in their "observations" as Cluster 3, the overall tone is pedagogical and concerned, rather than mockery. The general "caution" against the viral internet trend being an event with no consequences is present in this group, who appear to be analyzing the event as a sign of some broader, concerning socio-political trend.

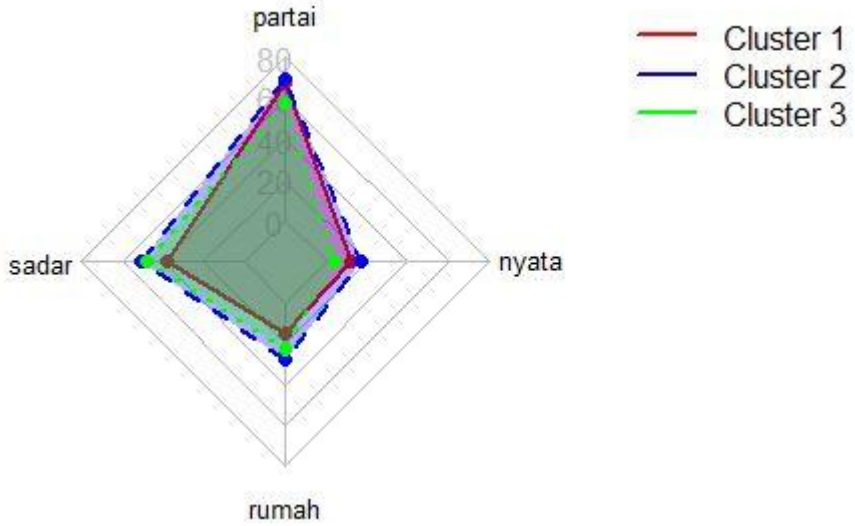

Figure 5. Strong words of Cluster 2 Cautious observers in radar plot

3.4.3. Cluster 3: The Political Protesters (n = 4)

In stark contrast to Cluster 1 and 2, this group is packed with a dense concentration of extreme, politically and emotionally charged words. Words such as lawan (fight/oppose), gibran (Vice-Presidential Gibran Rakabuming Raka), NKRI (Unitary State of the Republic of Indonesia), asset (commonly used for corrupted state officials in criminal cases), suara (voice/vote), muak (disgusted),

susah (hardship), kebebasan (freedom), and pilih (to choose/vote) form a coherent theme of political protest and direct opposition to the government. This cluster is extremely straightforward in its naming, and is one of "The Political Protesters". The general sentiment is extremely direct and clearly reflects disgust (muak) with economic hardship (susah) and calls for direct political action (lawan, pilih). The One Piece flag has clearly been adopted as an unambiguous banner of protest by this group, who used the opportunity to critique the government and political actors, name drop specific political figures, and express a sense of a great need for political change and freedom. For the members of this group, the flag has been stripped of all its anime-related context, and was co-opted as a tool of resistance against the state.

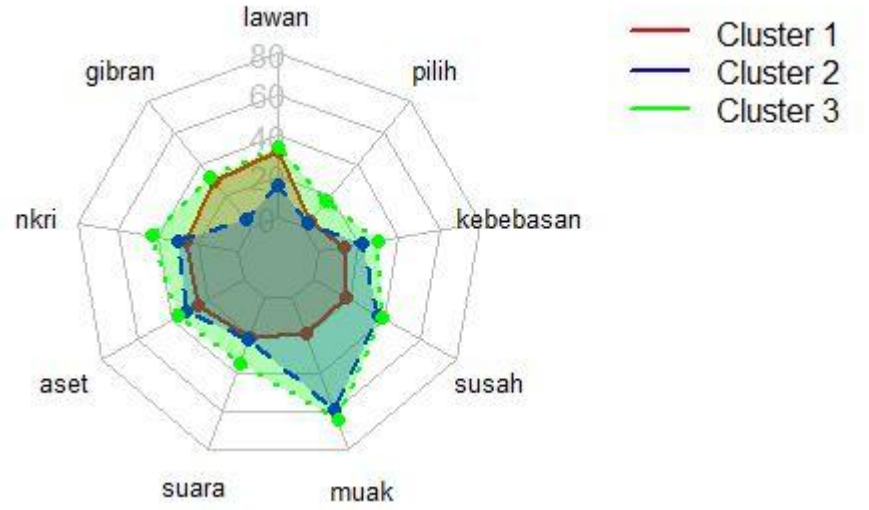

Figure 6. Strong words of Cluster 3 Political protesters in radar plot

This simple yet clear three-way segmentation illuminates the nature of the public interaction with this viral event in a very effective way. A single cultural artifact (simply flying a One Piece flag in public) is, simultaneously, interpreted in a very distinct way by three distinct groups as: (1) an apolitical piece of fun by Pop-Culture Enthusiasts, (2) a larger socio-political metaphor by Cautious Observers, and (3) a direct tool of political opposition and protest by Political Protesters. This confirms that the viral event was not a monolithic political movement, but a complex cultural moment in which a global pop cultural symbol was locally appropriated to various, very different ends, by audience members with very different levels of political engagement and dispositions within Indonesian society.

3.5. Implications

The carefully constructed, robust computational statistics pipeline has demonstrated that, far from being a monolith, the audience of this viral internet trend is distinctly segmented into three separate, statistically valid groups: The Pop-Culture Enthusiasts, The Cautious Observers, and The Political Protesters. This segmentation, as a result of this data-driven, empirical approach to understanding the subject, effectively maps the complex, multifaceted engagement of the public with the trend. It also unambiguously confirms that a single object (a One Piece flag) can be simultaneously both a tool of entertainment, cautious socio-political commentary, and outright political dissent. The public did not consume social media trend monolithically but interpreted it through cultural meaning [31].

While earlier work described it as a "potent symbol of protest and social aspiration" [8], the current results provide a nuanced corroboration and expansion of this statement. The existence of Pop-Culture Enthusiasts confirms that the global anime story is still at the forefront for the large proportion of the public, and, for this group, the local political meaning has been effectively "decoupled" from the symbol. The presence of Cautious Observers as a larger group of analytical, concerned (aware)

observers is also extremely informative about the process of how global cultural objects are locally appropriated in this specific cultural-historical moment.

From a methodological application, this research contributed to the validation as an integrated pipeline for high-dimensional text analysis prior to a statistical test. The application of PCA followed by medoid-based clustering and modified SigClust offers a different approach to some common practices which stop at internal validation indices [32]. By using PAM with Squared Euclidean distance, this work made it possible to make a more informed, customized distance choice as opposed to the default k-means implementation. The use of the modified SigClust as well as multiple-testing corrections introduced an additional layer of validation to the pipeline, beyond internal metrics (like silhouette and MSV scores) to a formal statistical test. It detected unsignificant a pair of clusters based on covariances [22] where analysis of variance for each variable [33] did not take covariances into account. This pipeline was not only effective for the present case but also highly replicable.

The methods is applicable outside of the 'Bendera One Piece' case study. A broad range of problems involving high-dimensional, low-sample-size (n << p) text data from social media like X (twitter) can be readily managed by the analytical pipeline. It is especially well-suited for tasks such measuring public participation in health initiatives, mapping discursive patterns in political campaigns, or tracking the development of brand perception. The key steps of feature selection (PCA) and clustering (medoid-based) with statistical validation are directly transferable.

3.6. Limitations and Future Work

This study used a lexicon-based preprocessing method and a Bag-of-Words model to identify themes in user comments. While this method worked well for our goal of video segmentation, it does not capture the deeper meanings that modern Natural Language Understanding models like BERT can offer. A useful direction for future research would be to look into using transformer-based embeddings for clustering. Instead of focusing solely on keyword frequency, this method may display more in-depth sub-segments according to sentiment or rhetorical style. However, this would also increase computational costs and create difficulties in interpreting the clustering results.

4. Conclusion

The goal of this exploratory study was to segment YouTube videos on a controversial socio-political topic of public interest, i.e. the viral phenomenon “Bendera One-Piece” in Indonesia, into subtopics grounded in user commentary, comments, and text from video descriptions. This so-called “unsupervised learning” problem on the common textbook front of clustering was characterized by the substantial high-dimensionality of the available source data (n << p). The integrated approach, comprising(1) a thorough text pre-processing and feature selection step to clean the document-term matrix, (2) dimensionality reduction via principal component analysis (PCA) to remedy the curse of dimensionality and extract the dominant latent thematic pattern, which was unambiguously socio-political in nature, (3) medoid-based (PAM) clustering with Squared Euclidean distance to compute an appropriate partition of the videos, successfully identified a coherent three-cluster solution. The interpretation of these validated clusters were (1) a non-political expression of fandom by Pop-Culture Enthusiasts, (2) a socio-political metaphor by Cautious Observers, and (3) a direct tool for political protest by Political Protesters.

In an important methodological extension of standard clustering, the three-part partition found by the medoid-based method was further subjected to a modified SigClust framework to obtain statistical evidence that the substructure imposed by these labels is present in the data and not a function of random noise. This was accomplished by treating the input data of the clustering algorithm as a set of (high-dimensional) features, and the medoid-based cluster labels as responses in Holm-adjusted pairwise significance tests, yielding p-values associated with the null hypothesis of a single underlying cluster.

In sum, this work delivers an interpretable analysis of a high-dimensional social media dataset, linking computational statistics and cultural inquiry with a descriptive patterning and inferential segmentation of online public discourse.

References

[1] S. Shajari and N. Agarwal, "Developing a network-centric approach for anomalous behavior detection on youtube," *Soc. Netw. Anal. Min.*, vol. 15, no. 3, 2025, doi: doi.org/10.1007/s13278-025-01417-y.

[2] P. Colás-Bravo and I. Quintero-Rodríguez, "YouTube as a Digital Resource for Sustainable Education," *Sustainability*, vol. 15, p. 5687, 2023, doi: doi.org/10.3390/su15075687.

[3] A. Shoufan and F. Mohamed, "YouTube and Education: A Scoping Review," *IEEE Access*, vol. 10, pp. 125576–125599, 2022, doi: doi: 10.1109/ACCESS.2022.3225419.

[4] M. E. Onder and O. Zengin, "YouTube as a source of information on gout: a quality analysis," *Rheumatol. Int.*, vol. 41, pp. 1321–1328, 2021, doi: doi.org/10.1007/s00296-021-04813-7.

[5] A. Finlayson, "YouTube and Political Ideologies: Technology, Populism and Rhetorical Form," *Polit. Stud.*, vol. 70, no. 1, pp. 62–80, 2022, doi: doi.org/10.1177/0032321720934630.

[6] D. B. V. Kaye and J. E. Gray, "Copyright Gossip: Exploring Copyright Opinions, Theories, and Strategies on YouTube," *Soc. Media Soc.*, vol. 7, no. 3, 2021, doi: doi.org/10.1177/20563051211036940.

[7] N. P. Nugroho, A. Faiz, D. Shabrina, and A. Antara, "Mengapa Pengibaran Bendera One Piece Dianggap Ancaman," *TEMPO*, Aug. 05, 2025. Accessed: Aug. 21, 2025. [Online]. Available: https://www.tempo.co/politik/mengapa-pengibaran-bendera-one-piece-dianggap-ancaman-2055142

[8] R. Benedetta, "Deciphering the World with One Piece: The power of manga in the study of international politics," Master Thesis, Ca' Foscari University of Venice, Italy, 2023. [Online]. Available: https://unitesi.unive.it/retrieve/0e99c6d0-ab50-439f-80d2-432cde748811/893705-1287503.pdf

[9] A. K. La'eng and H. Rosli, "Cultural Narratives and Global Impact: Analysing the Sociocultural Factorsof One Piece(1999) in Japanese Manga," *Res. Manag. Technol. Bus.*, vol. 5, no. 2, pp. 90–97, 2024, doi: doi.org/10.30880/rmtb.2024.05.02.012.

[10] M. Steinbach, L. Ertöz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in *New Directions in Statistical Physics*, Berlin: Springer, 2024. [Online]. Available: doi.org/10.1007/978-3-662-08968-2_16

[11] R. Vandaele, B. Kang, T. De Bie, and Y. Saeys, "The Curse Revisited: When are Distances Informative for the Ground Truth in Noisy High-Dimensional Data?," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022, pp. 2158–2172. [Online]. Available: https://proceedings.mlr.press/v151/vandaele22a.html

[12] S. Shajari, M. Alassad, and N. Agarwal, "Characterizing Suspicious Commenter Behaviors," *ASONAM 23 Proc. Int. Conf. Adv. Soc. Netw. Anal. Min.*, pp. 631–635, 2023, doi: doi.org/10.1145/3625007.3627309.

[13] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: doi.org/10.3390/app13063915.

[14] M. Farhan, R. D. L. R. Manik, H. R. Jannah, and L. H. Suadaa, "Comparison of Naive Bayes, K-Nearest Neighbor, and Support Vector Machine Classification Methods in SemiSupervised Learning for Sentiment Analysis of Kereta Cepat Jakarta Bandung (KCJB)," *Proc. 2023 Int. Conf. Data Sci. Off. Stat. ICDSOS*, vol. 1, 2023.

[15] M. A. Gandhi, S. P. Tripathy, S. S. Pawale, and J. S. Bhawalkar, "A narrative review with a step-by-step guide to R software for clinicians: Navigating medical data analysis in cancer research," *Cancer Res. Stat. Treat.*, vol. 7, no. 1, pp. 91–99, 2024, doi: doi.org/10.4103/crst.crst_313_23.

[16] B. Jeong and K. J. Lee, "NLP-Based Recommendation Approach for Diverse Service Generation," *IEEE Access*, vol. 12, 2024.

[17] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLOS ONE*, vol. 19, no. 12, p. e0315195, 2024, doi: doi.org/10.1371/journal.pone.0254937.

[18] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem, "Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling," *IEEE Access*, vol. 10, 2022.

[19] N. Oskolkov, "Dimensionality Reduction: Overview, Technical Details, and Some Applications," in *Applied Data Science in Tourism*, Switzerland: Springer, 2022, pp. 151–167. [Online]. Available: doi.org/10.1007/978-3-030-88389-8

[20] W. Budiaji and F. Leisch, "Simple K-Medoids Partitioning Algorithm for Mixed Variable Data," *Algorithms*, vol. 12, no. 9, p. 177, 2019, doi: https://doi.org/10.3390/a12090177.

[21] W. Budiaji, "Medoid-based Shadow Value Validation and Visualization," *Int. J. Adv. Intell. Inform.*, vol. 5, no. 2, pp. 76–88, 2019, doi: https://doi.org/10.26555/ijain.v5i2.326.

[22] Y. Liu, D. N. Hayes, A. Nobel, and J. S. Marron, "Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data," *J. Am. Stat. Assoc.*, vol. 103, no. 483, pp. 1281–1293, 2008.

[23] P. K. Kimes, Y. Liu, D. N. Hayes, and J. S. Marron, "Statistical significance for hierarchical clustering," *Biometrics*, vol. 73, no. 3, pp. 811–821, 2017.
[24] G. Sood, *tuber: Access YouTube from R*. (2023). [Online]. Available: https://CRAN.R-project.org/package=tuber
[25] K. Benoit, D. Muhr, and K. Watanabe, *stopwords: Multilingual Stopword Lists*. (2021). [Online]. Available: https://CRAN.R-project.org/package=stopwords
[26] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *J. Soft Comput. Data Min.*, vol. 2, no. 1, 2021.
[27] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, *cluster: Cluster Analysis Basics and Extensions*. (2023). [Online]. Available: https://CRAN.R-project.org/package=cluster
[28] W. Budiaji, *kmed: Distance-Based k-Medoids*. (2022). [Online]. Available: https://CRAN.R-project.org/package=kmed
[29] C. E. Agbangba, E. S. Aide, H. Honfo, and R. G. Kakai, "On the use of post-hoc tests in environmental and biological sciences: A critical review," *Heliyon*, vol. 10, no. 3, p. e25131, 2024, doi: doi.org/10.1016/j.heliyon.2024.e25131.
[30] A. Dolgun and H. Demirhan, "Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution," *Commun. Stat. - Simul. Comput.*, vol. 46, no. 7, pp. 5166–5183, 2017, doi: doi.org/10.1080/03610918.2016.1146761.
[31] J. Li, H. M. Adnan, and J. Gong, "Exploring Cultural Meaning Construction in Social Media: An Analysis of Liziqi's YouTube Channel," *J. Intercult. Commun.*, vol. 23, no. 4, 2023.
[32] S. Tripathi *et al.*, "Evaluation of Clustering with PCA for Market Segmentation: A Study Using Simulated and Surrogate Data," *Procedia Comput. Sci.*, vol. 253, pp. 2063–2075, 2025.
[33] S. Consuegra-Jiménez, C. Tovio-Gracia, and R. Vivas-Reyes, "Unsupervised learning techniques for clustering analysis of physicochemical properties in the periodic table elements," *Results Chem.*, vol. 17.