JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Clustering YouTube Comments on Mental Health in Indonesia Using the K-Means Algorithm

Agung Nugroho*, Raissa Amanda Putri

Fakultas Sains Dan Teknologi, Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
Email: [1*]agung0701222060@uinsu.ac.id, [2]raissa.ap@uinsu.ac.id,
Email Penulis Korespondensi: agung0701222060@uinsu.ac.id
Submitted **29-01-2026;** Accepted **06-04-2026;** Published **28-04-2026**

Abstract

This study aims to analyze mental health expressions in Indonesian-language YouTube comments using a text mining approach and the K-Means clustering algorithm. The increasing use of social media as a platform for expressing psychological conditions has resulted in large volumes of unstructured textual data that are difficult to analyze manually. Therefore, this study applies text preprocessing techniques, including case folding, tokenization, stopword removal, and stemming, followed by Term Frequency–Inverse Document Frequency (TF-IDF) weighting to transform textual data into numerical representations. The clustering process is performed using the K-Means algorithm, and the optimal number of clusters is determined using the Elbow Method and Silhouette Coefficient. The results show that the optimal number of clusters is k = 3, with the highest Silhouette Coefficient value indicating good cluster quality. A total of 2,411 YouTube comments were successfully grouped into three clusters, representing different types of mental health expressions, namely complaint expressions, personal experience narratives, and general responses. This study contributes by providing a social media comment clustering model to analyze mental health expressions in the Indonesian digital context. The results demonstrate that the K-Means algorithm can effectively identify meaningful patterns in large-scale textual data without requiring labeled datasets, making it useful for supporting data-driven mental health analysis.

Keywords: Mental Health; YouTube Comments; Text Mining; K-Means Clustering; TF-IDF

1. INTRODUCTION

The rapid development of information and communication technology has driven the growth of social media as a highly active space for public interaction. Social media is not only used as a medium for sharing information and entertainment, but also serves as a platform for users to express their emotional and psychological states. One of the platforms with a high level of user participation is YouTube. Through its comment feature, users can openly convey opinions, personal experiences, and emotional conditions regarding various issues, including mental health concerns [1]. This phenomenon positions YouTube comments as a potential source of textual data that can be analyzed to understand the psychological conditions of society [2].

Mental health is defined as a state of well-being in which individuals are able to realize their potential, cope with daily life stress, work productively, and contribute to their community [3]. In recent years, mental health problems have increased significantly, particularly among adolescents and university students. Academic pressure, social demands, and excessive exposure to digital information have been identified as factors that exacerbate individuals' psychological conditions [4], [5]. However, public awareness and utilization of formal mental health services remain relatively low, leading many individuals to express emotional distress through social media platforms [6].

YouTube has become one of the platforms frequently used to discuss mental health topics through educational content, personal experiences, and motivational videos. The comment sections of such videos often contain expressions of anxiety, stress, emotional exhaustion, and even social support among users [7]. When systematically analyzed, these comments can provide insights into emotional and psychological patterns within society. Nevertheless, the large volume of comments and the unstructured nature of textual data make manual analysis inefficient and highly subjective [8].

Text mining and *Natural Language Processing* (NLP) approaches can be employed as solutions to automatically process and analyze large-scale textual data [9]. Text mining enables information extraction, pattern recognition, and text grouping based on specific characteristics. One widely used technique in unlabeled text analysis is *clustering*, which groups data based on similarity without requiring labeled training data (*unsupervised learning*) [10].

The K-Means algorithm is a popular clustering method due to its conceptual simplicity, computational efficiency, and ability to handle large datasets [11]. In text analysis, K-Means is commonly combined with weighting methods such as *Term Frequency–Inverse Document Frequency* (TF-IDF) to transform textual data into numerical representations [12]. This combination allows YouTube comments to be grouped into clusters that share similar themes or psychological contexts.

Several previous studies have applied K-Means in social media data analysis. Ahmad *et al.* [13] utilized K-Means to cluster public opinions on Twitter and successfully identified dominant sentiment patterns. Another study by Sari and Wibowo [14] applied text mining techniques to social media data for mental health issue analysis, but focused primarily on positive and negative sentiment classification. Putra *et al.* [15] employed machine learning approaches to detect depression from social media text; however, the methods used were supervised, requiring labeled datasets.

Subsequent research by Rahmawati *et al.* [16] implemented K-Means for clustering health-related textual data, but the data source was not derived from social media platforms. Meanwhile, Hidayat and Nugroho [17] analyzed YouTube

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

comments related to social issues without specifically linking them to mental health aspects. Furthermore, several studies have not applied comprehensive cluster quality evaluation methods to determine the optimal number of clusters [18].

Based on the review of previous studies, several important limitations can be identified. Most prior research focused primarily on sentiment classification (positive, negative, and neutral), which does not fully capture the complexity of mental health expressions in social media data [14], [15]. In addition, some studies employed supervised learning approaches that require labeled datasets, limiting their scalability when dealing with large volumes of unstructured data [15]. Although K-Means has been widely used for clustering, its application in analyzing Indonesian-language YouTube comments specifically related to mental health remains limited [16], [17]. Furthermore, previous studies have rarely incorporated comprehensive cluster evaluation methods, such as the Elbow Method and Silhouette Coefficient, to determine the optimal number of clusters [18].

Therefore, this study aims to analyze Indonesian-language YouTube comments related to mental health using a text mining approach and the K-Means clustering algorithm. The study also applies TF-IDF weighting and cluster evaluation techniques, including the Elbow Method and Silhouette Coefficient, to ensure optimal and reliable clustering results.

The contribution of this research is threefold. First, it provides a clustering-based model for analyzing mental health expressions using Indonesian-language social media data. Second, it demonstrates the effectiveness of unsupervised learning in identifying meaningful patterns without requiring labeled datasets. Third, it integrates cluster evaluation methods to improve the validity and interpretability of the clustering results.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study employs a quantitative approach using text mining methods and clustering techniques to analyze mental health expressions in Indonesian-language YouTube comments. The research stages are systematically designed to ensure that data processing, method implementation, and result evaluation are conducted in a structured manner and produce outputs aligned with the research objectives. The overall research workflow in clustering Indonesian-language YouTube comments on mental health is illustrated in Figure 1.

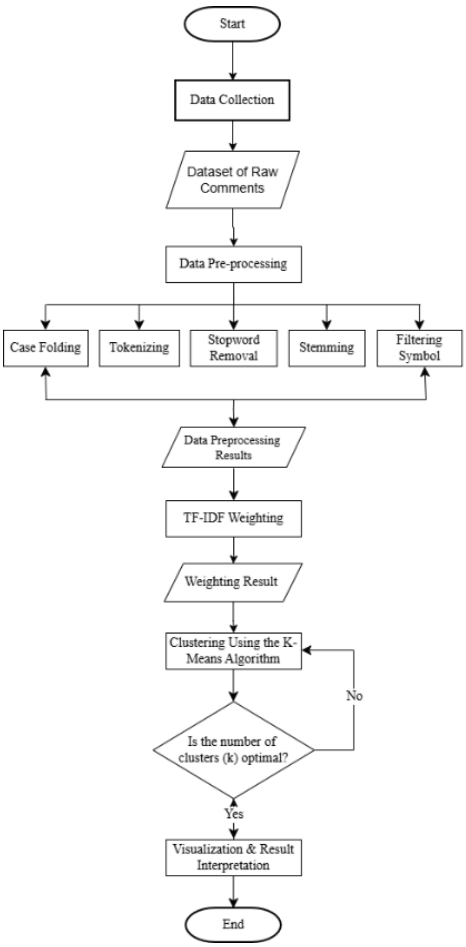

Figure 1. Clustering of Indonesian YouTube Comments on Mental Health Using the K-Means Algorithm

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Based on Figure 1, the research process begins with data collection from YouTube comments, followed by text preprocessing stages such as case folding, tokenization, stopword removal, stemming, and filtering. The processed data are then transformed into numerical representations using TF-IDF weighting. Furthermore, the K-Means algorithm is applied to cluster the data, and the optimal number of clusters is determined using evaluation methods. The final stage involves visualization and interpretation of the clustering results to identify patterns of mental health expressions.

The first stage is data collection. The data were obtained from public comments on a YouTube video discussing stress and overthinking published by the *1% Indonesian Life School* channel. Data collection was carried out using a web scraping technique with the assistance of the *YouTube Comment Downloader* library in the Google Colab environment. A total of 2,411 comments were collected and stored in a spreadsheet format. Each comment consists of several attributes, including username, comment text, upload date, and number of likes. In this study, the comment text attribute was used as the primary data for analysis. A summary of the dataset characteristics is presented in Table 1.

Table 1. Dataset Characteristics

Attribute	Description
Data source	YouTube comments
Platform	YouTube
Number of comments	2,411
Language	Indonesian
Data period	2023
Data collection method	Web scraping
Primary attribute	Comment text
Supporting attributes	Username, upload date, number of likes
Research focus	Mental health expressions

The second stage is text preprocessing, which aims to clean and prepare textual data for further analysis. The preprocessing steps include case folding to convert all text into lowercase letters, removal of punctuation, numbers, URLs, and special characters, tokenization to split text into individual words, stopword removal, and stemming to reduce words to their root forms. This process is essential to reduce noise and lexical variations, thereby improving the quality of the clustering results. The sequence of preprocessing steps follows the workflow shown in Figure 1.

The third stage is feature extraction using the Term Frequency–Inverse Document Frequency (TF-IDF) method. At this stage, the preprocessed text data are transformed into numerical representations. Term Frequency (TF) measures the frequency of a word within a document, while Inverse Document Frequency (IDF) reduces the weight of words that frequently appear across many documents. The combination of TF and IDF produces TF-IDF weights that represent the importance of each term within a document. The resulting TF-IDF matrix serves as the input for the clustering process.

The fourth stage is clustering using the K-Means algorithm. K-Means groups comments based on similarity by calculating the distance between data points and cluster centroids. In this stage, several values of the number of clusters (k) are tested to obtain the most optimal clustering results.

The fifth stage is clustering evaluation. This evaluation aims to determine the optimal number of clusters and assess the quality of the clustering results. The evaluation methods used are the Elbow Method and the Silhouette Coefficient. The Elbow Method analyzes changes in the Within-Cluster Sum of Squares (WCSS), while the Silhouette Coefficient measures the degree of cohesion within clusters and separation between clusters.

The final stage is result interpretation. In this stage, each cluster is analyzed based on dominant terms and the characteristics of the comments within it. The interpretation results are used to identify patterns of mental health expressions in YouTube comments, including clusters representing psychological complaints, clusters of personal experience narratives, and clusters of general responses or reactions to the content and other users' comments.

2.2 Text Mining Method and K-Means Algorithm

The problem-solving approach in this study employs a text mining framework to process unstructured YouTube comment data. One of the main stages in text mining is term weighting using the Term Frequency–Inverse Document Frequency (TF-IDF) method, which aims to represent textual data in numerical form so that it can be processed by clustering algorithms [3].

Term Frequency (TF) indicates the frequency of occurrence of a term in a document relative to the total number of words in that document. Mathematically, TF is calculated as follows:

$$TF(t,d) = \frac{number\ of\ occurrences\ of\ term\ t}{total\ number\ of\ therms\ in\ document\ d} \quad (1)$$

As an example, in the third document containing the text *"ngerasa capek"* with a total of two words, each word appears once. Therefore, the Term Frequency values for the words *"ngerasa"* and *"capek"* are both 0.5. Based on Table 2, each term in the document has a TF value of 0.5, indicating that both words appear with equal frequency. This shows that each term contributes equally to the representation of the document in the TF weighting process. To illustrate the calculation of Term Frequency (TF), an example is presented in Table 2.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Table 2. Example of TF Calculation

Kata	TF
ngerasa	0.5
capek	0.5

The final TF-IDF weight is obtained by multiplying the TF value by the Inverse Document Frequency (IDF) value. The TF-IDF calculation is formulated as follows:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (2)$$

Based on Table 3, the TF-IDF values indicate that the term "capek" has a higher weight compared to "ngerasa". This shows that the term "capek" is more significant in representing the content of the document, as it appears less frequently across documents but is more distinctive within the specific document. To illustrate the TF-IDF calculation process, an example is presented in Table 3.

Table 3. Example of TF-IDF Calculation

Kata	TF	IDF	TF-IDF
ngerasa	0.5	0.916	0.458
capek	0.5	1.609	0.805

To present the TF-IDF weighting results in a structured manner, a TF-IDF matrix is constructed as shown in Table 4. In this matrix, each row represents a document, while each column represents a term or feature. The values in the matrix indicate the TF-IDF weight of each term in a specific document.

Table 4. Example of TF-IDF Matrix

Dokumen	ngerasa	capek	makasih	video	overthinking	tidur
D1	0	0	0	0	0	0
D2	0	0	0	0	0	0
D3	0.458	0.805	0	0	0	0
D4	0	0	0	0	0.536	0.536
D5	0.183	0	0.321	0.321	0	0

Based on Table 4, it can be observed that each row represents a document and each column represents a term. The TF-IDF values indicate the importance of terms within specific documents, where higher values reflect stronger relevance in representing the document content. This matrix serves as the main input for the K-Means algorithm to cluster comments based on similarity in word patterns and mental health expressions.

After the TF-IDF weighting process, the data are clustered using the K-Means algorithm [11]. This algorithm groups data into a predefined number of clusters based on the shortest distance between data points and cluster centroids [4]. The clustering results are then evaluated using the Elbow Method and the Silhouette Coefficient to determine the optimal number of clusters and assess clustering quality.

The K-Means algorithm operates by specifying the number of clusters (k) [11] and assigning data points to clusters based on the minimum distance to the cluster centroids. Initially, centroids are randomly selected, and distances between each comment vector and the centroids are calculated. Each data point is assigned to the nearest cluster. The centroids are then updated based on the mean position of all data points within each cluster. This iterative process continues until the clustering results stabilize and no significant changes occur in the centroid positions.

In this study, the distance between data points and centroids is measured using Euclidean Distance [11], which is commonly applied in K-Means due to its simplicity and computational efficiency in high-dimensional feature spaces such as TF-IDF matrices. The use of Euclidean Distance enables clustering of comments based on similarities in word usage patterns that represent users' emotional expressions.

To determine the optimal number of clusters, the Elbow Method and the Silhouette Coefficient are employed. The Elbow Method is used to observe changes in the Within-Cluster Sum of Squares (WCSS) [5] across different numbers of clusters. The optimal number of clusters is identified at the point where the decrease in WCSS begins to level off, indicating a balance between the number of clusters and clustering quality.

In addition, the Silhouette Coefficient is used to evaluate the quality of the formed clusters [5] by measuring the degree of cohesion within clusters and separation between clusters. Higher Silhouette Coefficient values indicate better-defined clusters with clearer separation. The combination of these two evaluation methods ensures that the clustering results obtained in this study are optimal and reliable.

Based on the evaluation results, the number of clusters used in this study is three. The clustering results are further analyzed to identify the characteristics and patterns of mental health expressions in YouTube comments, which are discussed in detail in the Results and Discussion section.

3. RESULT AND DISCUSSION

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

This section presents the results of implementing the text mining approach and the K-Means clustering algorithm in analyzing Indonesian-language YouTube comments related to mental health. The results include the evaluation of the optimal number of clusters using the Elbow Method and the Silhouette Coefficient, as well as quantitative analysis of clustering performance. Furthermore, the results are interpreted based on the characteristics of comments in each cluster. All experiments were conducted using the Python programming language with the support of text mining and machine learning libraries.

3.1 Results of Data Processing and Evaluation

This subsection presents the quantitative results of clustering evaluation, including the determination of the optimal number of clusters and the assessment of clustering quality. The data processing and evaluation stage aims to assess the performance of the K-Means algorithm in clustering YouTube comments based on TF-IDF feature representations. The evaluation focuses on determining the optimal number of clusters and assessing the quality of the resulting clustering structure.

a. Implementation of Text Mining and K-Means Clustering
The implementation of the text mining process begins with preprocessing the collected YouTube comment data. At this stage, several steps are carried out, including case folding, tokenization, stopword removal, stemming, and filtering, to clean and standardize the textual data. The purpose of this process is to reduce noise and ensure that the data are suitable for further analysis. After preprocessing, the cleaned text data are transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) method. This transformation produces a TF-IDF matrix that represents the importance of each term within each document. Furthermore, the TF-IDF matrix is used as input for the K-Means clustering algorithm. The clustering process is carried out by determining the number of clusters (k) and calculating the distance between data points and cluster centroids using Euclidean Distance. Each data point is assigned to the nearest cluster, and the centroid positions are updated iteratively until convergence is achieved. The results of this process produce clusters of YouTube comments based on similarities in word patterns, which represent different types of mental health expressions. These clustering results are then evaluated to determine the optimal number of clusters and to assess the quality of the clustering structure.

b. Determination of the Optimal Number of Clusters Using the Elbow Method
The Elbow Method is employed to determine the optimal number of clusters by analyzing changes in the Within-Cluster Sum of Squares (WCSS) across different numbers of clusters. WCSS represents the total squared distance between each data point and the cluster centroid within the same cluster. A smaller WCSS value indicates higher homogeneity within clusters. In this study, experiments were conducted by varying the number of clusters from $k = 2$ to $k = 6$. The resulting WCSS values were visualized using an Elbow graph. The graph shows a significant decrease in WCSS values up to $k = 3$, followed by a gradual flattening for higher values of k. To determine the optimal number of clusters, the Elbow Method is applied by analyzing the WCSS values for different numbers of clusters, as illustrated in Figure 2.

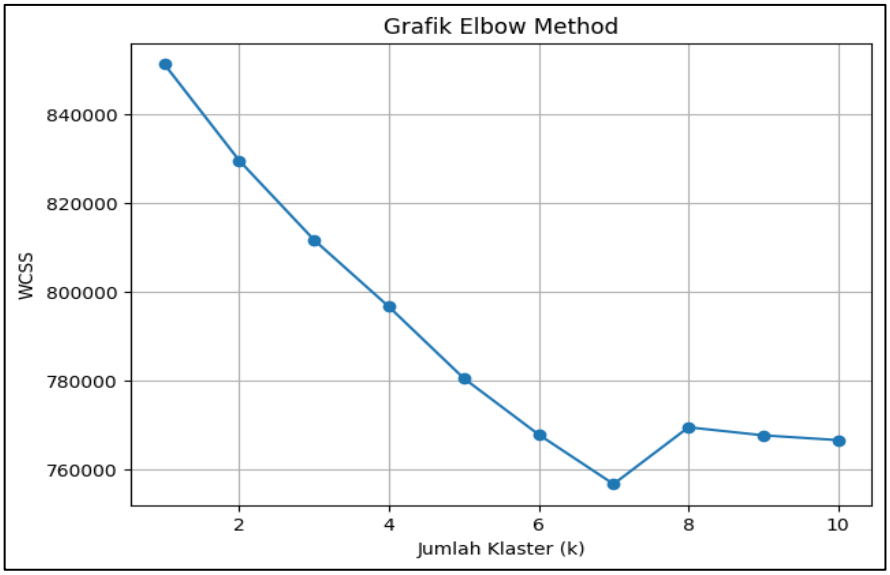

Figure 2. Elbow Method Graph

Based on Figure 2, the WCSS value decreases significantly as the number of clusters increases from k = 2 to k = 3. After k = 3, the decrease in WCSS becomes more gradual, indicating that adding more clusters does not provide a significant improvement in data grouping. Therefore, the elbow point is identified at k = 3, which represents the optimal number of clusters for this study.

c. Cluster Quality Evaluation Using the Silhouette Coefficient

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

In addition to the Elbow Method, cluster quality evaluation was also performed using the Silhouette Coefficient. The Silhouette Coefficient measures how well a data point fits within its assigned cluster compared to other clusters. The coefficient value ranges from -1 to 1, where values closer to 1 indicate better cluster assignment and clearer separation between clusters. The Silhouette evaluation was conducted using the same range of cluster values as in the Elbow analysis, namely $k = 2$ to $k = 6$. The results show that the highest average Silhouette Coefficient value was obtained at $k = 3$. To further evaluate the clustering quality, the Silhouette Coefficient is used to measure the cohesion and separation of clusters, as shown in Figure 3.

```
[ ]
from sklearn.metrics import silhouette_score

silhouette_scores = []

for k in range(2, 7):
    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(X)
    score = silhouette_score(X, labels)
    silhouette_scores.append(score)

# Menampilkan hasil silhouette
for k, score in zip(range(2,7), silhouette_scores):
    print(f'k = {k}, Silhouette Score = {score:.4f}')

...
k = 2, Silhouette Score = 0.6860
k = 3, Silhouette Score = 0.4468
k = 4, Silhouette Score = 0.5216
k = 5, Silhouette Score = 0.5176
k = 6, Silhouette Score = 0.4126
```

Figure 3. Silhouette Coefficient Graph

Based on Figure 3, the Silhouette Coefficient reaches its highest value at k = 3 compared to other cluster values. This indicates that the clustering structure at k = 3 has better cohesion within clusters and clearer separation between clusters. Therefore, the results of the Silhouette evaluation confirm that three clusters represent the optimal configuration for this study.

d. Implementation of the K-Means Algorithm Using Python

After determining the optimal number of clusters, the clustering process was performed using the K-Means algorithm with $k = 3$. The implementation was carried out using the Python programming language with the *scikit-learn* library. The comment data, which had been represented in the form of a TF-IDF matrix, served as the primary input for the clustering process.

The K-Means algorithm operates by randomly initializing cluster centroids and then assigning data points to the nearest centroid using the Euclidean Distance metric. The algorithm iteratively updates the centroid positions based on the mean of data points within each cluster until convergence is achieved and no significant changes in centroid positions occur.

The clustering results show that all YouTube comments were successfully grouped into three clusters based on the similarity of word patterns. Each cluster represents a different type of mental health expression, indicating that the K-Means algorithm is capable of identifying meaningful patterns within the data. These results provide a strong foundation for further interpretation and discussion in the subsequent section.

3.2 Discussion

This discussion section focuses on interpreting the clustering results and explaining the meaning of each cluster in the context of mental health expressions found in YouTube comments. The analysis is conducted by examining the characteristics of the comments and the dominant terms within each cluster.

a. Cluster 0 – Complaints

Cluster 0 is dominated by comments containing expressions of severe mental conditions, such as stress, emotional exhaustion, anxiety, cognitive pressure, and overthinking. Comments in this cluster generally take the form of emotional venting, reflecting the psychological state of individuals experiencing high emotional burdens or mental instability. Users in this cluster tend to use words that describe mental fatigue and life pressure, whether related to work, social relationships, or personal problems. This pattern indicates that the YouTube comment section is not only used for light interaction, but also serves as a space for expressing negative emotions and emotional release. The existence of this complaints cluster suggests that social media, particularly YouTube, has the potential to serve as an early indicator of public mental health conditions. This finding is consistent with previous studies that state social media platforms can reflect users' psychological states through language use and emotional expressions.

b. Cluster 1 – Narratives

Cluster 1 consists of narrative comments in which users share personal experiences, thought processes, and life journeys in dealing with mental health issues. Comments in this cluster are generally longer and delivered in a reflective manner, without exhibiting extreme emotional expressions. Narrative comments often contain elements of introspection and learning from personal experiences. Users not only express their difficulties, but also describe how they attempt to understand and cope with their mental conditions. This indicates a higher level of self-awareness regarding mental health.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9552
Hal 537-544
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

This cluster reflects the role of YouTube as a platform for sharing experiences and providing indirect social support. Users utilize the comment section to share their stories with the hope of being understood or offering value to others who may be experiencing similar conditions.

c. Cluster 2 – Responses

Cluster 2 includes comments that function as general responses to the video, such as expressing agreement, relating to the video content, providing brief opinions, or simply reacting to the presented material. Comments in this cluster are generally short, surface-level, and not focused on personal experiences.

This cluster contains the largest number of comments compared to the other clusters. This reflects the interaction pattern of the majority of YouTube users, who tend to participate in discussions in a reactive and concise manner. Users are more likely to provide quick responses rather than express in-depth personal experiences or mental conditions.

Although relatively simple, the responses cluster still plays an important role in the analysis, as it indicates the level of audience engagement with mental health-related content. This cluster also emphasizes that not all social media interactions represent deep psychological conditions.

d. Interpretation and Comparison with Previous Studies

The clustering results demonstrate that an unsupervised learning approach is capable of grouping YouTube comments based on mental health expression patterns without requiring prior labeling. This represents a key advantage over classification methods that rely on labeled datasets.

Compared to previous studies that employ sentiment analysis with positive, neutral, and negative categories, this study provides a more contextual perspective. The resulting clusters reflect not only emotional polarity, but also the types of expressions used by users, such as complaints, personal narratives, and brief responses.

These findings reinforce prior research suggesting that text mining and clustering are effective approaches for analyzing large-scale social media data. Furthermore, the focus on Indonesian-language YouTube comments contributes additional value to digital mental health research, which has predominantly focused on English-language data.

4. CONCLUSION

This study applies a text mining approach using TF-IDF weighting and the K-Means clustering algorithm to analyze Indonesian-language YouTube comments related to mental health. The evaluation using the Elbow Method and Silhouette Coefficient indicates that the optimal number of clusters is three. The clustering results reveal distinct patterns of mental health expressions, including complaint expressions, personal experience sharing, and general responses. The Silhouette Coefficient value shows that the clustering structure has good cohesion and separation between clusters. These findings demonstrate that the proposed approach is effective in identifying patterns of mental health expressions from unstructured social media data. However, this study is limited to data collected from a single YouTube video. Future research is recommended to expand the dataset and compare multiple clustering methods to obtain more comprehensive results.

REFERENCES

[1] S. P. Aji, N. Ani, And R. Mar'atu Sholihah, "Faktor-Faktor Yang Berpengaruh Terhadap Kondisi Kesehatan Mental Mahasiswa Pada Proses Pembelajaran Daring Di Masa Pandemi Covid-19 Factors That Influence The Students' Mental Health Conditions On The Online Learning Processes In The Covid-19 Pandemic," Jurnal Ilmu Kesehatan Masyarakat Berkala, Vol. 4, No. 1, Pp. 28–37, 2022.

[2] S. P. Aji, N. Ani, And R. Mar'atu Sholihah, "Faktor-Faktor Yang Berpengaruh Terhadap Kondisi Kesehatan Mental Mahasiswa Pada Proses Pembelajaran Daring Di Masa Pandemi Covid-19 Factors That Influence The Students' Mental Health Conditions On The Online Learning Processes In The Covid-19 Pandemic," Jurnal Ilmu Kesehatan Masyarakat Berkala, Vol. 4, No. 1, Pp. 28–37, 2020.

[3] E. Annuril Akbar, W. Astuti, J. J. Sakai, And N. Y. Aulia, "Indonesian Journal Of Digital Public Relations (Ijdpr) Analisis Konten Kesehatan Mental Akun Youtube Cnn Indonesia Mental Health Content Analysis Of Cnn Indonesia Youtube Account," 2024. [Online]. Available: Https://Journals.Telkomuniversity.Ac.Id/Ijdpr

[4] D. Ariel And T. Handayani, "Jurnal Ilmu Komputer Dan Sistem Informasi Perbandingan Efektivitas Algoritma K-Means Da N Fuzzy C-Means Untuk Clustering Data Produksi Alpukat Di Indonesia," 2025.

[5] A. Atira And B. Nurina Sari, "Penerapan Silhouette Coefficient, Elbow Method Dan Gap Statistics Untuk Penentuan Cluster Optimum Dalam Pengelompokkan Provinsi Di Indonesia Berdasarkan Indeks Kebahagiaan," Jurnal Ilmiah Wahana Pendidikan, Vol. 9, No. 17, Pp. 76–86, 2023, Doi: 10.5281/Zenodo.8282638.

[6] Aulia And Anggi, "Gambaran Kesehatan Mental Mahasiswa Di Masa Pandemi Covid-19," Online, 2021. [Online]. Available: Http://Ejurnalmalahayati.Ac.Id/Index.Php/Duniakesmas/Index

[7] E. Ayuningsih, S. R. Lubis, And Z. Tembusai, "Jurnal Media Informatika Budidarma Perancangan Ai Chatbot (Stylesavvy) Untuk Memilih Fashion Pakaian Berdasarkan Warna Kulit Di Toko Xi-Xiu," Vol. 8, Pp. 1790–1794, 2024, Doi: 10.30865/Mib.V8i3.8220.

[8] S. A. Azzahra And N. W. A. Majid, "Klasifikasi Dan Analisis Semantik Cyberbullying Sosial Media X: Integrasi Web Scraping Dan Natural Language Processing (Nlp)," Jurnal Educatio Fkip Unma, Vol. 11, No. 2, Apr. 2025, Doi: 10.31949/Educatio.V11i2.12725.

[9] H. Tetiawadi, P. R. Studi Manajemen Informatika Politeknik Malinau Jl Ladang, D. Malinau Seberang, K. Malinau Utara Kabupaten Malinau, And K. Utara, "Sistem Informasi Publik Sekretariat Dprd Kabupaten Malinau," Jurnal Bangkit Indonesia, Vol. 12, No. 01, 2023.

[10] Bella Salsa Risnawati, Nasichah Nasichah, Muhammad Faqih Prayogo, And Zannuby Al Izzami, "Faktor-Faktor Yang Mempengaruhi Kesehatan Mental Mahasiswa Bimbingan Dan Penyuluhan Islam Uin Syarif Hidayatullah Jakarta," Jurnal Ilmiah Dan Karya Mahasiswa, Vol. 2, No. 1, Pp. 179–186, Dec. 2023, Doi: 10.54066/Jikma.V2i1.1389.
[11] S. Bila And R. Sharafi, "Identifikasi Pola Diskusi Publik Mengenai Pemindahan Ibu Kota Negara Menggunakan Analisis Tf-Idf Dan K-Means Clustering," Seminar Nasional Sistem Informasi, 2024.
[12] F. Bintang Putra, M. Taufik Chulkamdi, And F. Febrinita, "Implementasi Data Mining Untuk Memprediksi Data Stok Fukubi Outfit Menggunakan Metode K-Nearest Neighbor," 2024.
[13] T. A. Br Sembiring And M. S. Hasibuan, "Text Clustering In Karo Language Using Tf-Idf Weighting And K-Means Clustering," Jurnal Teknik Informatika (Jutif), Vol. 4, No. 5, Pp. 1257–1265, Nov. 2023, Doi: 10.52436/1.Jutif.2023.4.5.1462.
[14] S. Budi And H. Sakur, "Analisis Perbandingan Pengukuran Jarak Pada Algoritme K-Means Berbasis Sum Of Square Error," 2022.
[15] S. Chancellor And M. De Choudhury, "Methods In Predictive Techniques For Mental Health Status On Social Media: A Critical Review," Dec. 01, 2020, Nature Research. Doi: 10.1038/S41746-020-0233-7.
[16] A. R. Danurisa And J. Heikal, "Customer Clustering Using The K-Means Clustering Algorithm In The Top 5 Online Marketplaces In Indonesia," 2023, Doi: 10.33258/Birci.V5i3.6450.
[17] G. Erda, C. Gunawan, And Z. Erda, "Grouping Of Poverty In Indonesia Using K-Means With Silhouette Coefficient," Parameter: Journal Of Statistics, Vol. 3, No. 1, Pp. 1–6, Jun. 2023, Doi: 10.22487/27765660.2023.V3.I1.16435.
[18] A. Fauzan, "Mental Dalam Youtube Channel Satu Persen Skripsi," 2023.
[19] G. Risky Pratiwi, D. Wahiddin, E. E. Awal, A. Fauzi, U. Buana, And P. Karawang, "Klasterisasi Tingkat Kemiskinan Kabupaten/Kota Di Indonesia Menggunakan Algoritma K-Means Dan K-Medoids," 2024, Doi: 10.33364/Algoritma/V.21-2.1788.
[20] Febby Wilyani, Qonaah Nuryan Arif, And Fitri Aslimar, "Pengenalan Dasar Pemrograman Python Dengan Google Colaboratory," Jurnal Pelayanan Dan Pengabdian Masyarakat Indonesia, Vol. 3, No. 1, Pp. 08–14, Mar. 2024, Doi: 10.55606/Jppmi.V3i1.1087.