JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Pengelompokkan Film Trending di Youtube Menggunakan TF-IDF dan K-Means Clastering

Dwi Remawati[1], Hendro Wijayanto[2], Yustina Retno Wahyu Utami[3], Bayu Dwi Raharja[4]
[1,4]Teknologi Informasi, STMIK Sinar Nusantara Surakarta
[2,3]Informatika, STMIK Sinar Nusantara Surakarta
Email: [1]dwirema@sinus.ac.id, [2]hendro@sinus.ac.id, [3]yustina.lecturer@sinus.ac.id, [4]bayudr@sinus.ac.id
Email Penulis Korespondensi: dwirema@sinus.ac.id

Abstrak
YouTube telah menjadi platform utama untuk konsumsi konten video, dengan tren video yang terus berkembang sesuai perubahan minat audiens. Penelitian ini bertujuan untuk mengelompokkan film trending di YouTube berdasarkan judul dan popularitasnya menggunakan pendekatan TF-IDF dan K-Means Clustering. TF-IDF digunakan untuk mengekstraksi fitur dari judul video, mengidentifikasi kata-kata kunci penting yang mencirikan tema setiap film. Algoritma K-Means kemudian digunakan untuk mengelompokkan video ke dalam beberapa cluster berdasarkan kemiripan fitur TF-IDF dan jumlah views. Hasil penelitian menunjukkan bahwa video dapat dikelompokkan ke dalam tiga cluster dengan karakteristik unik. Cluster pertama berisi video dengan tema komedi, aktor populer, dan jumlah views tinggi. Cluster kedua mencakup video dari berbagai genre dengan jumlah views yang bervariasi. Cluster ketiga terdiri dari video yang lebih spesifik dengan popularitas tinggi. Evaluasi menggunakan Silhouette Score menunjukkan bahwa kualitas clustering masih dapat ditingkatkan. Penelitian ini memberikan wawasan bagi kreator konten dan pemasar digital untuk menyusun strategi konten yang lebih menarik dan relevan, serta memberikan kontribusi akademis dalam analisis data berbasis teks.
Kata Kunci: data mining, clustering, K-means, TF IDF, film trending

Abstract
YouTube has become a major platform for video content consumption, with video trends constantly evolving according to changing audience interests. This study aims to cluster trending movies on YouTube based on their titles and popularity using TF-IDF and K-Means Clustering approaches. TF-IDF is used to extract features from video titles, identifying important keywords that characterize the theme of each movie. The K-Means algorithm is then used to cluster videos into several clusters based on the similarity of TF-IDF features and the number of views. The results show that videos can be grouped into three clusters with unique characteristics. The first cluster contains videos with comedy themes, popular actors, and high views. The second cluster includes videos from various genres with varying views. The third cluster consists of more specific videos with high popularity. Evaluation using Silhouette Score shows that the quality of clustering can still be improved. This study provides insights for content creators and digital marketers to develop more interesting and relevant content strategies, as well as providing academic contributions in text-based data analysis.
Keywords*: data mining, clustering, K-means, TF IDF, film trending*

1. PENDAHULUAN

Di era digital saat ini, YouTube telah menjadi salah satu media utama bagi pengguna untuk menikmati konten hiburan, termasuk film[1]. Setiap hari, banyak orang menonton dan mencari film-film trending pada platform ini. Dengan banyaknya konten yang tersedia, pengelompokkan film berdasarkan judul dan popularitasnya menjadi sangat penting[2][3]. Dengan adanya pengelompokan ini dapat membantu pengguna, kreator, dan pengiklan untuk lebih memahami pola preferensi audiens serta tren populer dalam berbagai kategori film. Pengelompokkan film berdasarkan judul dan popularitas (views) memiliki manfaat yang signifikan, terutama dalam analisis data dan industri hiburan[4][5]. Manfaatnya antara lain membantu memahami pilihan penonton, mengoptimalkan strategi konten dan pemasaran, serta menyederhanakan proses rekomendasi yang dapat membantu pengguna menemukan konten yang mungkin mereka sukai berdasarkan film trending yang memiliki genre atau kategori serupa[6][7]. Selain itu, hasil pengelompokan ini mendukung pengambilan keputusan dalam industri hiburan, khususnya dalam menentukan jenis film yang sebaiknya diproduksi atau didistribusikan, sehingga meningkatkan peluang keberhasilan komersial. Bagi pengiklan, pengelompokan ini membuka peluang untuk menjangkau audiens yang lebih spesifik. Sebagai contoh, pengiklan produk olahraga mungkin akan lebih tertarik beriklan pada film aksi atau petualangan yang sedang trending, karena segmen audiensnya lebih relevan.

Klusterisasi (*clustering*) adalah teknik dalam data mining atau machine learning yang berfungsi untuk membagi data menjadi beberapa kelompok (cluster) berdasarkan kesamaan atribut atau karakteristik tertentu[8]. Data dalam satu cluster akan memiliki kemiripan yang tinggi, sementara data antar cluster akan memiliki perbedaan yang signifikan. Penggunaan TF-IDF (*Term Frequency-Inverse Document Frequency*) dalam clustering film bertujuan untuk mengekstraksi informasi penting dari teks, yaitu judul film, dan mengubahnya menjadi fitur numerik yang dapat dianalisis dalam proses clustering[9]. TF-IDF menghitung seberapa penting suatu kata dalam dokumen (dalam hal ini, judul film) pada seluruh kumpulan dokumen[10]. Metode ini memprioritaskan kata-kata yang spesifik dan signifikan dalam judul film, seperti Action, Comedy, atau Romantic, yang dapat menggambarkan kategori atau tema film tersebut. Komponen IDF (*Inverse Document Frequency*) dalam TF-IDF berfungsi untuk menurunkan nilai kata-kata yang terlalu umum dan sering muncul di banyak dokumen, seperti “Movie” atau “Film”. Dengan demikian kata-kata yang kurang informatif dalam konteks

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

clustering tidak akan memberikan pengaruh besar pada hasil analisis kemiripan antar judul[8]. Clustering membutuhkan input berupa data numerik, dan TF-IDF menghasilkan vektor yang mengandung bobot setiap kata penting[11][12]. Representasi ini memungkinkan setiap judul film diterjemahkankan ke dalam ruang fitur numerik. Kata-kata dengan bobot TF-IDF yang tinggi mencerminkan tema atau genre spesifik, sehingga dapat membantu dalam proses pengelompokan film berdasarkan tema atau kategori. TF-IDF adalah metode yang sederhana namun efisien untuk diterapkan pada data teks[13]. Meskipun tidak sekompleks teknik NLP lainnya, TF-IDF cukup kuat untuk mengidentifikasi pola-pola dasar dalam teks tanpa membutuhkan sumber daya komputasi yang besar. Ketika digunakan bersama algoritma clustering seperti K-Means, TF-IDF memberikan representasi numerik yang memungkinkan K-Means mengelompokkan film dengan lebih akurat berdasarkan tema atau popularitas[14][12]. Hasilnya, film dengan tema atau genre serupa dapat dikelompokkan bersama, menciptakan cluster yang bermakna berdasarkan pola kunci dalam judul film.

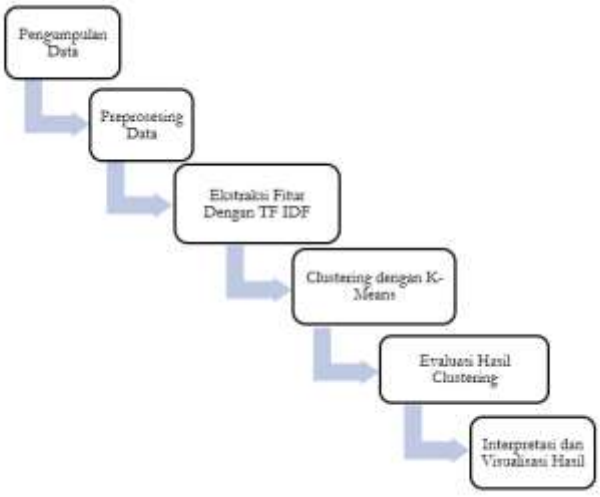
Salah satu penelitian yang telah dilakukan oleh [15] membahas bagaimana dokumen skripsi dikelompokkan berdasarkan topiknya untuk membentuk kelompok-kelompok topik skripsi. Setelah dikelompokkan, data divisualisasikan berdasarkan periode waktu tahun untuk menganalisis trend topik skripsi yang berkembang sehingga dapat digunakan sebagai referensi mahasiswa dalam memilih topik skripsi yang akan diambil dan untuk membantu pembimbing dalam menyetujui topik penelitian skripsi yang baru. Untuk pengelompokkan menggunakan text mining dan K-Means Clustering. Tujuannya untuk melakukan analisis trend untuk mengetahui trend topik skripsi. Penelitian berikutnya dilakukan oleh [16] mengelompokkan berita dengan tujuan untuk mengidentifikasi setiap kelompok berita. Menggunakan metode K-Means yang didasarkan pada proses pembobotan kata menggunakan Algoritma TF-IDF (*Term Frequency Inverse Document Frequency*). Proses clustering menggunakan berita hasil crawling dari situs detik.com dengan jangka waktu satu tahun (2018) yang berjumlah 124.509 berita dan disimpan dalam bentuk file CSV (*Comma Seperated Value*). Berdasarkan hasil pengujian, terdapat 27 kelompok berita yang berhasil di identifikasi dengan kemampuan aplikasi yang cukup memadai dalam memproses data yang besar. Penelitian berikutnya melakukan pegelompokkan review konsumen terhadap salah satu produk skincare. Produk skincare tersebut adalah SKINTIFIC merupakan merk lokal Indonesia yang sedang viral dengan penjualan yang cukup banyak baik offline maupun online. Pada penelitian ini review diambil dari twitter, menggunakan metode TF- IDF untuk pembobotan kata dan metode K-Means Clustering untuk pengelompokkan sentimen positif dan negatif. Hasil penelitian, pengelompokkan dibagi menjadi 2 yaitu Cluster Positif dan Cluster negatif. Dengan persentase jumlah cluster positif (C1) 79,5%. Sedangkan persentase jumlah cluster negatif (C2) 19,7% [17].

Pada penelitian ini diharapkan menghasilkan pengelompokan film berdasarkan judul dan popularitas di platform YouTube. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi efektivitas algoritma clustering dalam mengelompokkan data video berdasarkan judul dan popularitas, sehingga dapat menjadi dasar untuk pengembangan model analisis data yang lebih baik di masa depan.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian dalam proses clustering menggunakan algoritma TF-IDF dan metode K-mean untuk pengelompokan film trending di youtube meliputi beberapa tahapan yaitu mulai dari penentuan data, preprosesing data yang terdiri dari tokenisasi, stemming, penghilangan stopword serta eksplorasi data menggunakan bahasa pemrograman Python. Tahapan penelitian pada penelitian ini seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Penjelasan tahapan penelitian :

1. Pengumpulan Data
   Dataset yang digunakan adalah data trending youtube movies yang diperoleh secara online melalui Kaggle dengan alamat situs https://www.kaggle.com/datasets.
2. Preprosesing Data, pada tahap ini untuk memudahkan penerapan preprosesing data, penulis menggunakan nltk dalam Phyton.
   a. Pembersihan Teks: Judul film akan dibersihkan dari karakter khusus, angka, dan kata-kata yang tidak relevan (stopwords) untuk meningkatkan kualitas pengelompokan. Proses ini mencakup case folding, tokenisasi, stopword removal dan stemming.
   b. Penormalan Atribut Popularitas: Data numerik seperti jumlah views, likes, dan komentar akan dinormalisasi menggunakan teknik Min-Max Scaling untuk menyamakan skala antara judul teks dan popularitas dalam analisis berikutnya.
3. Ekstraksi Fitur Menggunakan TF-IDF
   a. Setiap judul film yang telah melalui pra-pemrosesan akan diubah menjadi representasi numerik menggunakan TF-IDF. TF-IDF digunakan untuk menghitung bobot setiap kata dalam judul film, dengan memperhatikan frekuensi kemunculan kata tersebut di seluruh judul yang ada, sehingga kata-kata yang lebih khas dari setiap judul akan mendapatkan bobot lebih tinggi.
   b. Hasil ekstraksi fitur ini akan digunakan sebagai representasi vektor dari masing-masing judul untuk proses pengelompokan.

   Komponen utama TF-IDF adalah:
   - Term Frequency (TF): Mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Formula umumnya:

$$TF\ (t,d) = \frac{jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{Total\ kata\ dalam\ dokumen\ d}$$

   - Inverse Document Frequency (IDF): Mengukur seberapa unik sebuah kata dalam keseluruhan dokumen di corpus. Formula umumnya:

$$IDF(t,D) = log\ \left(\frac{N}{1+nt}\right)$$

     - N: Jumlah total dokumen dalam corpus.
     - nt : Jumlah dokumen yang mengandung kata t.
     - Penambahan 1 pada penyebut untuk menghindari pembagian dengan nol.
   - TF-IDF Score: Menggabungkan TF dan IDF untuk memberikan bobot pada sebuah kata dalam dokumen:

$$TF{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

4. Clustering dengan K-Means
   a. Algoritma K-Means Clustering akan digunakan untuk mengelompokkan data berdasarkan fitur TF-IDF dari judul dan atribut popularitas.
   b. Proses clustering ini bertujuan untuk membentuk kelompok-kelompok yang menggambarkan tren film trending di YouTube, baik dari sisi tema (berdasarkan judul) maupun dari tingkat popularitasnya.
   c. Jarak Antar Vektor (Cosine Similarity atau Euclidean Distance):
   Cosine Similarity sering digunakan untuk data teks karena lebih fokus pada arah vektor daripada magnitudenya:

$$d(xi,cj) = 1-\cos(\theta) = 1-\frac{xi.cj}{\|xi\|\|cj\|}$$

   - xi: Vektor TF-IDF dari dokumen ke-i.
   - cj: Centroid cluster ke-j.
   - ‖xi‖: Panjang (norm) dari vektor xi.

   d. Euclidean Distance juga bisa digunakan untuk mengukur jarak:

$$d(xi,cj) = \sqrt{\sum_{i=1}^{N}(xik - cjk)2}$$

   e. Fungsi Objektif : Sama seperti K-Means standar, fungsi objektif adalah meminimalkan jarak total antara dokumen dan centroid cluster.

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

$$j = \sum_{j=1}^{k} \sum_{\{x_i \in C_j\}} d(xi - cj)^2$$

5. Evaluasi Hasil Clustering
   a. Hasil clustering akan dievaluasi menggunakan Silhouette Score untuk memastikan bahwa kluster yang terbentuk memiliki kualitas dan keterpisahan yang baik.
   b. Analisis hasil kluster akan dilakukan dengan membandingkan karakteristik setiap kelompok yang terbentuk, termasuk perbedaan dalam atribut popularitas dan kata kunci yang dominan dalam judul di setiap kluster.
6. Interpretasi dan Visualisasi Hasil
   a. Hasil kluster akan divisualisasikan dalam bentuk scatter plot untuk menunjukkan distribusi film berdasarkan kategori yang terbentuk. Visualisasi ini membantu dalam memahami perbedaan tema atau tren film di setiap kelompok.
   b. Berdasarkan hasil klustering, akan diinterpretasikan karakteristik masing-masing kelompok, seperti apakah ada tema khusus yang mendominasi kluster tertentu atau apakah ada kluster yang didominasi oleh film-film dengan popularitas

3. HASIL DAN PEMBAHASAN

Clustering film trending di Youtube ini dimulai dari pengumpulan data training set, tahap desain, dan penerapan data mining. Nantinya, akan dihasilkan pola tertentu yang dapat digunakan untuk melihat dan memprediksi tren. Tahapan penelitian meliputi identifikasi dat training set, pra-processing data, eksplorasi data menggunakan python, dan analisis hasil.

3.1 Data set

Data yang diperoleh dari Kaggle merupakan data mentah. Gambar 2 menampilkan data mentah sebelum dilakukan konversi pada kolom Youtube Views. Pada kolom Youtube Views, data terlihat dalam bentuk string, misalnya "9.5M views", "188K views", dan sebagainya.

Data Awal (15 Baris Pertama):

Movies Title	Youtube Channel	Youtube Views
Fryday (2018) - Full Movie - Superhit Comedy Movie | Govinda Sanjay Mishra Varun Sharma	Bollywood Premium	9.5M views
Kumar Sanu Romantic Song || Best of Kumar Sanu Duet Super Hit 90's Songs Old Is Gold Song 2024	Jeet Music	188K views
Maayon - New Released South Indian Hindi Dubbed Movie 2024 | South Dubbed Movie | South Movie 2024	Wamindia Movies	24M views
Behen Hogi Teri (2017) Full Movie 4K | Rom-Com Movie | Rajkummar Rao | Shruti Haasan | Gautam Gulati	Bollywood Premium	15M views
Shahid Kapoor - Batti Gul Meter Chalu (2018) Shraddha Kapoor | Latest Bollywood Release	2000s Ki Filmein	35M views
Shaadi Mein Zaroor Aana (2017) Full Hindi Movie (4K) Rajkumar Rao Kriti Kharbanda	Ultra Movie Parlour	38M views
Top 10 Bollywood Comedy Scenes - Akshay Kumar - Paresh Rawal - Johnny Lever - Rajpal Yadav	Shemaroo Comedy	5.2M views
GREAT HACK - Blockbuster Hindi Dubbed Action Movie | Sree Vishnu Chitra Shukla | South Action Movie	Hindi Dubbed Movie Talkies	5.1M views
Khatta Meetha | Superhit Hindi Comedy Movie | Akshay Kumar - Johny Lever - Asrani - Rajpal Yadav	Shemaroo Comedy	78M views
Garam Masala (HD) Full Movie | Hindi Comedy Movies | Akshay Kumar Movies | Latest Bollywood Movies	Venus Entertainment	111M views
Top 10 Highest Imdb Rated Films 2022 ??? #shorts #imdb	FILMY CRUSH	2.3M views
New Released South Indian Hindi Dubbed Movie 2024 | New 2024 Hindi Dubbed Action Movie| The Real Don	South Prime Cinema	20M views
Genius 2018 Full Movie (4K) Utkarsh Sharma Nawazuddin Siddiqui Ishitha Chauhan | Full Hindi Movie	Ultra Movie Parlour	25M views
Old Vs New Bollywood Mashup 2024 | Superhits Romantic Hindi Songs Mashup Live - DJ MaShUP 2024	New Hindi PartyMix	2.4K views
10 Highest Grossing Movies | Top 10 Highest Grossing Indian movies of All Time	Factonic	1.1M views

Gambar 2. Data mentah

3.2 Data Preprocessing

Bagian ini berisi tahapan penyiapan data, yang terdiri dari case folding, tokenisasi, stemming, dan penghilangan kata henti serta kata sambung. Penulis menggunakan bahasa pemrograman python pada semua tahapan penyiapan ini agar pengolahan data menjadi mudah dan efektif.

a. Case folding

Merupakan langkah dalam preprocessing data teks yang bertujuan untuk mengkonversi semua huruf dalam teks menjadi huruf kecil (lowercase). Langkah ini dilakukan untuk memastikan bahwa perbedaan kapitalisasi huruf tidak memengaruhi analisis teks. Hasil case folding pada judul-judul video atau film dari dataset seperti pada Gambar 3.

	Case Folding
0	fryday (2018) - full movie - superhit comedy m...
1	kumar sanu romantic song || best of kumar sanu...
2	maayon - new released south indian hindi dubbe...
3	behen hogi teri (2017) full movie 4k | rom-com...
4	shahid kapoor - batti gul meter chalu (2018) s...
5	shaadi mein zaroor aana (2017) full hindi movi...
6	top 10 bollywood comedy scenes - akshay kumar ...
7	great hack - blockbuster hindi dubbed action m...
8	khatta meetha | superhit hindi comedy movie |...
9	garam masala (hd) full movie | hindi comedy mo...
10	top 10 highest imdb rated films 2022 ??? #shor...
11	new released south indian hindi dubbed movie 2...
12	genius 2018 full movie (4k) utkarsh sharma naw...
13	old vs new bollywood mashup 2024 | superhits r...
14	10 highest grossing movies | top 10 highest gr...

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Gambar 3. Hasil Case Folding

b. Tokenisasi

Tokenisasi adalah proses dalam pengolahan bahasa alami (*Natural Language Processing* atau NLP) yang memecah teks menjadi unit-unit kecil yang disebut token. Token biasanya berupa kata, frasa, kalimat, atau simbol, tergantung pada konteks dan tujuan analisis. Hasil tokenisasi seperti pada Gambar 4.

	Tokenized
0	[fryday, (, 2018,), -, full, movie, -, superh...
1	[kumar, sanu, romantic, song, ||, best, of, ku...
2	[maayon, -, new, released, south, indian, hind...
3	[behen, hogi, teri, (, 2017,), full, movie, 4...
4	[shahid, kapoor, -, batti, gul, meter, chalu, ...
5	[shaadi, mein, zaroor, aana, (, 2017,), full,...
6	[top, 10, bollywood, comedy, scenes, -, akshay...
7	[great, hack, -, blockbuster, hindi, dubbed, a...
8	[khatta, meetha, |, superhit, hindi, comedy, m...
9	[garam, masala, (, hd,), full, movie, |, hind...
10	[top, 10, highest, imdb, rated, films, 2022, ?...
11	[new, released, south, indian, hindi, dubbed, ...
12	[genius, 2018, full, movie, (, 4k,), utkarsh,...
13	[old, vs, new, bollywood, mashup, 2024, |, sup...
14	[10, highest, grossing, movies, |, top, 10, hi...

Gambar 4. Hasil Tokenisasi

c. Stopwords Removal

Stopwords removal adalah proses dalam untuk menghapus kata-kata umum (*stopwords*) yang sering muncul dalam teks tetapi tidak memberikan informasi penting untuk analisis. Hasil stopwords removal ditunjukkan pada gambar 5.

Hasil Stopwords Removal (15 baris pertama):

	Stopwords Removed
0	[fryday, (, 2018,), -, full, movie, -, superh...
1	[kumar, sanu, romantic, song, ||, best, kumar,...
2	[maayon, -, new, released, south, indian, hind...
3	[behen, hogi, teri, (, 2017,), full, movie, 4...
4	[shahid, kapoor, -, batti, gul, meter, chalu, ...
5	[shaadi, mein, zaroor, aana, (, 2017,), full,...
6	[top, 10, bollywood, comedy, scenes, -, akshay...
7	[great, hack, -, blockbuster, hindi, dubbed, a...
8	[khatta, meetha, |, superhit, hindi, comedy, m...
9	[garam, masala, (, hd,), full, movie, |, hind...
10	[top, 10, highest, imdb, rated, films, 2022, ?...
11	[new, released, south, indian, hindi, dubbed, ...
12	[genius, 2018, full, movie, (, 4k,), utkarsh,...
13	[old, vs, new, bollywood, mashup, 2024, |, sup...
14	[10, highest, grossing, movies, |, top, 10, hi...

Gambar 5. Hasil Stopwords Removal

d. Stemming

Stemming merupakan proses untuk mengubah kata menjadi bentuk dasarnya (root word atau stem) dengan cara menghapus imbuhan seperti awalan, akhiran, atau sisipan. Proses stemming tidak selalu menghasilkan kata yang valid secara linguistik, tetapi cukup untuk tujuan analisis teks. Hasil stemming seperti pada Gambar 6.

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

```
Hasil Stemming (15 baris pertama):
                                              Stemmed
0   [fryday, (, 2018, ), -, full, movi, -, superhi...
1   [kumar, sanu, romant, song, ||, best, kumar, s...
2   [maayon, -, new, releas, south, indian, hindi,...
3   [behen, hogi, teri, (, 2017, ), full, movi, 4k...
4   [shahid, kapoor, -, batti, gul, meter, chalu, ...
5   [shaadi, mein, zaroor, aana, (, 2017, ), full,...
6   [top, 10, bollywood, comedi, scene, -, akshay,...
7   [great, hack, -, blockbust, hindi, dub, action...
8   [khatta, meetha, |, superhit, hindi, comedi, m...
9   [garam, masala, (, hd, ), full, movi, |, hindi...
10  [top, 10, highest, imdb, rate, film, 2022, ?, ...
11  [new, releas, south, indian, hindi, dub, movi,...
12  [genius, 2018, full, movi, (, 4k, ), utkarsh, ...
13  [old, vs, new, bollywood, mashup, 2024, |, sup...
14  [10, highest, gross, movi, |, top, 10, highest...
```

Gambar 6. Hasil Stemming

3.3 Penerapan TF IDF

Algoritma *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) untuk menghitung setiap token (kata) pada setiap dokumen dalam korpus. Semakin sering kata muncul, maka semakin besar pula nilai TF. Selanjutnya, mencari nilai IDF untuk menghitung seberapa banyak istilah-istilah tersebut tersebar luas dalam kumpulan dokumen terkait. Berbeda dengan TF, dalam IDF, semakin jarang kata-kata muncul dalam dokumen, semakin besar nilainya. Hasil TF IDF seperti pada gambar 6.

```
Hasil TF-IDF (15 baris pertama):
          10      2017      2018  2023      2024        4k    action  \
0   0.000000  0.000000  0.407854   0.0  0.000000  0.000000  0.000000
1   0.000000  0.000000  0.000000   0.0  0.128008  0.000000  0.000000
2   0.000000  0.000000  0.000000   0.0  0.373463  0.000000  0.000000
3   0.000000  0.407980  0.000000   0.0  0.000000  0.377203  0.000000
4   0.000000  0.000000  0.330480   0.0  0.000000  0.000000  0.000000
5   0.000000  0.458854  0.000000   0.0  0.000000  0.424238  0.000000
6   0.205017  0.000000  0.000000   0.0  0.000000  0.000000  0.000000
7   0.000000  0.000000  0.000000   0.0  0.000000  0.000000  0.581912
8   0.000000  0.000000  0.000000   0.0  0.000000  0.000000  0.000000
9   0.000000  0.000000  0.000000   0.0  0.000000  0.000000  0.000000
10  0.203733  0.000000  0.000000   0.0  0.000000  0.000000  0.000000
11  0.000000  0.000000  0.000000   0.0  0.382794  0.000000  0.218331
12  0.000000  0.000000  0.349168   0.0  0.000000  0.349168  0.000000
13  0.000000  0.000000  0.000000   0.0  0.269630  0.000000  0.000000
14  0.367668  0.000000  0.000000   0.0  0.000000  0.000000  0.000000

    adventur    akshay  allu  ...  underwat   utkarsh     varun  viral  \
0        0.0  0.000000   0.0  ...       0.0  0.000000  0.488036    0.0
1        0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
2        0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
3        0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
4        0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
5        0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
6        0.0  0.288480   0.0  ...       0.0  0.000000  0.000000    0.0
7        0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
8        0.0  0.344613   0.0  ...       0.0  0.000000  0.000000    0.0
9        0.0  0.339501   0.0  ...       0.0  0.000000  0.000000    0.0
10       0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
11       0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
12       0.0  0.000000   0.0  ...       0.0  0.417814  0.000000    0.0
13       0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
14       0.0  0.000000   0.0  ...       0.0  0.000000  0.000000    0.0
```

Gambar 7. Hasil TF IDF

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

3.4 Hasil klusterisasi

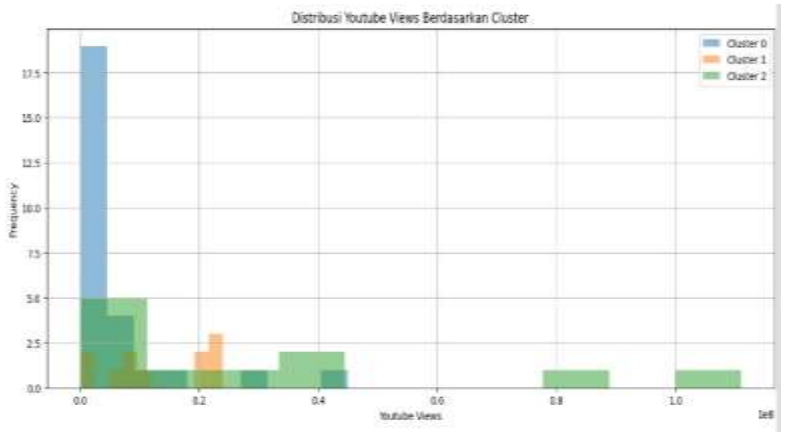
Hasil vektor TF-IDF sebagai input ke algoritma K-Means untuk klusterisasi atau pengelompokan. Hasil klusterisasi seperti pada Gambar 7, dimana terdapat 3 pengelompokkan berdasarkan Genre yaitu Comedy, Romantic dan lainnya. Algoritma K-Means mengelompokkan data film ini ke dalam beberapa cluster berdasarkan kemiripan atribut (seperti judul, views, kategori, atau atribut lain yang digunakan). Setiap angka di kolom "Cluster" menunjukkan label cluster yang dihasilkan oleh K-Means.Terdapat 3 kluster yaitu 0, 1 dan 2.

	Processed Title	Youtube Views	\
0	fryday (2018) - full movi - superhit comedi ...	9500000.0	
1	kumar sanu romant song || best kumar sanu duet...	188000.0	
2	maayon - new releas south indian hindi dub mov...	24000000.0	
3	behen hogi teri (2017) full movi 4k | rom-co...	15000000.0	
4	shahid kapoor - batti gul meter chalu (2018)...	35000000.0	
5	shaadi mein zaroor aana (2017) full hindi mo...	38000000.0	
6	top 10 bollywood comedi scene - akshay kumar -...	5200000.0	
7	great hack - blockbust hindi dub action movi |...	5100000.0	
8	khatta meetha | superhit hindi comedi movi | a...	78000000.0	
9	garam masala (hd) full movi | hindi comedi m...	111000000.0	
10	top 10 highest imdb rate film 2022 ? ? ? # sho...	2300000.0	
11	new releas south indian hindi dub movi 2024 | ...	20000000.0	
12	genius 2018 full movi (4k) utkarsh sharma na...	25000000.0	
13	old vs new bollywood mashup 2024 | superhit ro...	2400.0	
14	10 highest gross movi | top 10 highest gross i...	1100000.0	

	Genre	Cluster
0	Comedy	2
1	Romantic	1
2	Others	1
3	Others	2
4	Others	2
5	Others	2
6	Comedy	2
7	Action	0
8	Comedy	2
9	Comedy	2
10	Others	2
11	Action	1
12	Others	2
13	Romantic	1
14	Others	0

Gambar 8. Hasil Klusterisasi

Persebaran views seperti terlihat pada Gambar 8 menunjukkan bahwa Cluster 0 (Bar Biru): Mayoritas video dalam cluster ini memiliki jumlah views yang sangat rendah, mendekati nol. Sebagian kecil video dalam cluster ini memiliki views hingga sekitar 40 juta, tetapi jumlahnya jauh lebih sedikit dibandingkan views rendah. Cluster 1 (Bar Oranye): Cluster ini menunjukkan distribusi views yang lebih tersebar dibandingkan Cluster 0. Sebagian besar video memiliki views rendah hingga sedang (di bawah 40 juta), tetapi ada beberapa video dengan views di rentang yang lebih tinggi. Cluster 2 (Bar Hijau): Cluster ini memiliki beberapa video dengan jumlah views sangat tinggi, bahkan mendekati 100 juta. Sebagian besar video dalam cluster ini memiliki views menengah hingga tinggi, menunjukkan bahwa cluster ini mungkin berisi video populer.



Gambar 9. Distribusi Views

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

3.5 Evaluasi klusterisasi

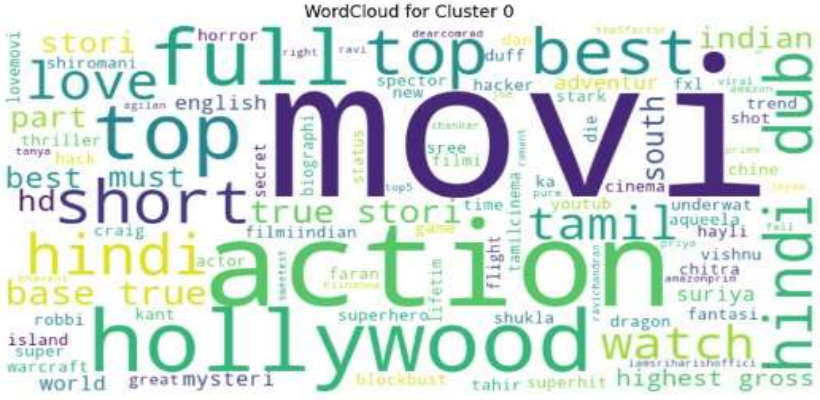
Silhouette adalah metrik yang digunakan untuk mengevaluasi kualitas hasil clustering. Metrik ini mengukur seberapa baik titik data cocok dengan cluster mereka sendiri dibandingkan dengan cluster lain. Pada penelitian ini setelah klusterisasi, Silhouette Score rata-rata yang diperoleh adalah 0,066358.

```
Silhouette Score (average): 0.06635818449762458
```

Dari hasil Silhouette Score rata-rata tersebut, menunjukkan kualitas clustering yang rendah. Titik data berada di batas antara cluster. Cluster yang terbentuk tidak cukup representatif untuk memisahkan data ke dalam kelompok yang jelas. Hal ini bisa dikarenakan representasi vektor yang dihasilkan memiliki dimensi tinggi, sehingga clustering menjadi sulit.

3.6 WordCloud

WordCloud adalah alat yang sering digunakan dalam analisis teks untuk memberikan gambaran tentang tema atau kata kunci yang sering digunakan dalam teks atau kumpulan data teks, dimana ukuran setiap kata menunjukkan frekuensi atau pentingnya kata tersebut. Pada penelitian ini, Cluster 0 pada Gambar 9 merepresentasikan video yang sebagian besar berkaitan dengan film aksi, baik dari Hollywood maupun industri film India, dengan tema-tema seperti daftar film terbaik, film berdasarkan kisah nyata, atau film populer.



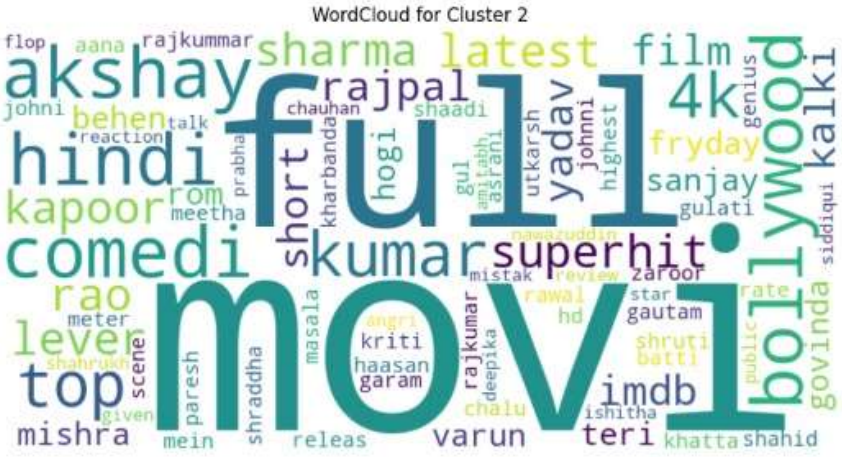
Gambar 10. WordCloud Cluster 0

Cluster 1 pada Gambar 10 merepresentasikan video yang sebagian besar berkaitan dengan Film berbahasa Hindi yang merupakan hasil dubbing dari bahasa lain, terutama dari industri film India Selatan. Konten populer yang mencakup film baru, mashup lagu, dan film blockbuster.

Gambar 11. WordCloud Cluster 1

Cluster 2 pada Gambar 11 merepresentasikan video yang berkaitan dengan Film Bollywood dengan fokus pada genre komedi. Film lengkap atau konten berkualitas tinggi (misalnya, 4K). Video yang menonjolkan aktor-aktor terkenal Bollywood, seperti Akshay Kumar dan Rajpal Yadav.

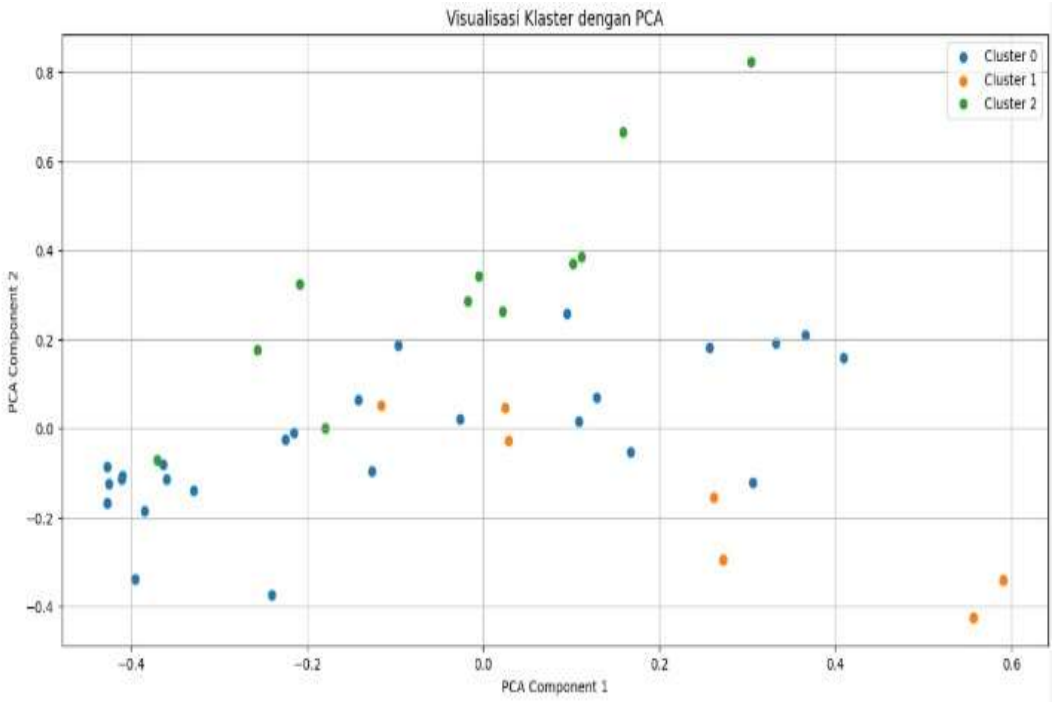
JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 1, Januari 2025, Hal 65-74
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi



Gambar 12. WordCloud Cluster 2

3.7 Visualisasi Kluster dengan PCA

Principal Component Analysis (PCA) adalah metode analisis data yang digunakan memvisualisasikan data, mengurangi dimensi dan memahami pola dalam dataset. PCA bekerja dengan mengubah dataset berdimensi tinggi menjadi ruang berdimensi lebih rendah sambil mempertahankan informasi sebanyak mungkin. Pada Gambar 12 merupakan visualisasi kluster dengan PCA pada kasus ini. Distribusi Klaster Cluster 0 (Biru): memiliki penyebaran data yang cukup luas dan dominan di sisi kiri dan tengah bawah grafik, Cluster 1 (Oranye): Data ini lebih tersebar di tengah grafik, tetapi dalam jumlah lebih kecil dibanding Cluster 0 dan Cluster 2 (Hijau): lebih tersebar di bagian atas dan kanan grafik, menunjukkan pola distribusi yang lebih spesifik.



Gambar 13. Visualisasi Kluster dengan PCA

4. KESIMPULAN

Penelitian ini berhasil mengelompokkan video YouTube menjadi tiga cluster utama berdasarkan fitur seperti judul video, genre, dan jumlah views. WordCloud untuk setiap cluster memberikan gambaran kata-kata dominan yang mencerminkan tema utama dalam cluster tersebut. Penelitian ini memberikan wawasan tentang bagaimana video YouTube dapat dikelompokkan berdasarkan tema dan popularitas menggunakan teknik TF-IDF, K-Means clustering, dan WordCloud. Hasil *Silhouette Score* rata-rata adalah sekitar 0.066, yang menunjukkan kualitas clustering yang rendah, menunjukkan bahwa data dalam cluster memiliki tumpang tindih yang tinggi, atau sulit untuk dipisahkan dengan baik ke

dalam cluster yang berbeda. Sebagai masukan untuk penelitian ke depan dengan optimalisasi jumlah klaster (k) ataupun dengan pemilihan fitur yang lebih relevan.

UCAPAN TERIMAKASIH

Ucapan terima kasih disampaikan kepada pihak-pihak yang telah mendukung terlaksananya penelitian ini yang tidak bisa penulis sebutkan satu per satu.

DAFTAR PUSTAKA

[1] D. Röchert, G. Neubaum, B. Ross, F. Brachten, and S. Stieglitz, "Opinion-based homogeneity on youtube: Combining sentiment and social network analysis," *Comput. Commun. Res.*, vol. 2, no. 1, pp. 81–108, 2020, doi: 10.5117/CCR2020.1.004.ROCH.

[2] E. D. Putra, M. H. Rifqo, D. Deslianti, and K. Krismiyani, "Analysis of The Theme Clustering Algorithm Using K-Means Method," *J. Komputer, Inf. dan Teknol.*, vol. 2, no. 2, pp. 431–442, 2022, doi: 10.53697/jkomitek.v2i2.884.

[3] S. Kasus, S. Ransomware, and D. Nasional, "Topic Modelling Berbasis Embedding pada Komentar YouTube," pp. 873–884, 2024.

[4] M. Riduwan, C. Fatichah, and A. Yuniarti, "Klasterisasi Dokumen Menggunakan Weighted K-Means Berdasarkan Relevansi Topik," *JUTI J. Ilm. Teknol. Inf.*, vol. 17, no. 2, p. 146, 2019, doi: 10.12962/j24068535.v17i2.a892.

[5] E. Susanto, V. C. Mawardi, and M. D. Lauro, "Aplikasi Clustering Berita Dengan Metode K Means Dan Peringkas Berita Dengan Metode Maximum Marginal Relevance," *J. Ilmu Komput. dan Sist. Inf.*, vol. 9, no. 1, p. 62, 2021, doi: 10.24912/jiksi.v9i1.11560.

[6] S. Krisdianto Sitanggang, F. Rakhmat Umbara, and H. Ashaury, "Klasifikasi Video Pada Media Sosial Youtube Dengan Menggunakan Metode K-Means Dan Support Vector Machine," *J. Locus Penelit. dan Pengabdi.*, vol. 2, no. 10, pp. 1027–1032, 2023, doi: 10.58344/locus.v2i10.1732.

[7] M. Rafi Haidar Arsyad, "Klusterisasi Data Review Pengguna Aplikasi Marketplace Blibli.Com dengan Algoritma K-Means dan K-Medoids," vol. 09, pp. 2657–1501, 2024.

[8] T. I. Saputra and R. Arianty, "Implementasi Algoritma K-Means Clustering Pada Analisis Sentimen Keluhan Pengguna Indosat," *J. Ilm. Inform. Komput.*, vol. 24, no. 3, pp. 191–198, 2019, doi: 10.35760/ik.2019.v24i3.2361.

[9] M. Darwis, G. T. Pranoto, Y. E. Wicaksana, and Y. Yaddarabullah, "Implementation of TF-IDF Algorithm and K-mean Clustering Method to Predict Words or Topics on Twitter," *JISA(Jurnal Inform. dan Sains)*, vol. 3, no. 2, pp. 49–55, 2020, doi: 10.31326/jisa.v3i2.831.

[10] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen," *Pros. Semin. Nas. Sist. Inf. dan Teknol.*, pp. 145–149, 2022.

[11] M. A. Haq, W. Purnomo, and N. Y. Setiawan, "Analisis Clustering Topik Survey menggunakan Algoritme K-Means (Studi Kasus: Kudata)," *... Teknol. Inf. dan Ilmu ...*, vol. 7, no. 7, pp. 3498–3506, 2023, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/13147%0Ahttps://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/13147/5928.

[12] I. W. Ardiyasa, "Penerapan K-Means Clustering untuk Klasifikasi Serangan Cyber pada Syslog File," *J. Sist. dan Inform.*, vol. 14, no. 2, pp. 143–149, 2020, doi: 10.30864/jsi.v14i2.305.

[13] A. Supriatman, "Pembobotan TF-IDF pada Judul Penelitian Dosen Sebagai Dasar Klasifikasi Menggunakan Algoritma K-NN (Studi Kasus: Universitas Siliwangi)," *J. Serambi Eng.*, vol. 6, no. 1, pp. 1573–1579, 2021, doi: 10.32672/jse.v6i1.2645.

[14] L. P. Refialy, H. Maitimu, and M. S. Pesulima, "Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster," *Techno.Com*, vol. 20, no. 2, pp. 321–329, 2021, doi: 10.33633/tc.v20i2.4572.

[15] M. R. Irianto, A. Maududie, and F. N. Arifin, "Implementation of K-Means Clustering Method for Trend Analysis of Thesis Topics (Case Study: Faculty of Computer Science, University of Jember)," *Berk. Sainstek*, vol. 10, no. 4, p. 210, 2022, doi: 10.19184/bst.v10i4.29524.

[16] S. Data *et al.*, "Clustering Berita Menggunakan Algoritma TF-IDF Dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.Com," vol. 3, no. 1, 2022.

[17] H. Nababan, I. Kelana Jaya, S. Manurung, and H. Artikel, "Analisis Sentimen Produk Penjualan Shopee Pada Pengguna Twitter Menggunakan Metode K-Means," *J. Ilm. Sist. Inf.*, vol. 3, no. 2, pp. 137–142, 2023, [Online]. Available: http://ojs.fikom-methodist.net/index.php/methosisfo.