Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Pengelompokan Analisis Sentimen Komentar *Youtube* Terhadap Pengambilalihan Jalan Rusak di Lampung Menggunakan Algoritma *Clustering*

Niko Purnomo[1*], Windu Gata[2]
Ilmu Komputer, Universitas Nusa Mandiri, Jakarta, Indonesia
*e-mail *Corresponding Author:* 14220038@nusamandiri.ac.id

Abstrak

Clustering is a method to group data into groups with certain similarities. This research analyzes text clustering on YouTube video comments about damaged road repairs in Lampung. Three clustering algorithms were used: K-means, DBSCAN, and HDBSCAN. The results showed a silhouette score for K-means of -0.348, DBSCAN of 0.836, and HDBSCAN of 0.106. Theme analysis on DBSCAN clusters showed better clustering than K-means and HDBSCAN. DBSCAN clusters are easier to infer because the topics of each cluster are well classified. Thus, DBSCAN proved superior in clustering text comments, with the highest silhouette score of 0.836% in the case of damaged road repair in Lampung.
Keywords: *Analisis Sentimen; Clustering; K-means; DBSCAN; HDBSCAN*

Abstrak

Clustering adalah metode untuk mengelompokkan data ke dalam kelompok-kelompok dengan kemiripan tertentu. Penelitian ini menganalisis pengelompokan teks pada komentar video YouTube tentang perbaikan jalan rusak di Lampung. Tiga algoritma *clustering* digunakan: K-means, DBSCAN, dan HDBSCAN. Hasil penelitian menunjukkan skor siluet untuk K-means sebesar -0,348, DBSCAN sebesar 0,836, dan HDBSCAN sebesar 0,106. Analisis tema pada cluster DBSCAN menunjukkan pengelompokan yang lebih baik dibandingkan K-means dan HDBSCAN. *Cluster* DBSCAN lebih mudah disimpulkan karena topik tiap *cluster* terklasifikasi dengan baik. Dengan demikian, DBSCAN terbukti lebih unggul dalam mengelompokkan komentar teks, dengan skor siluet tertinggi 0,836% pada kasus perbaikan jalan rusak di Lampung.
Kata Kunci*:* *Analisis Sentimen; Clustering, K-means; DBSCAN; HDBSCAN.*

1. Pendahuluan

Perbaikan jalan merupakan isu krusial dalam pembangunan infrastruktur, terutama di negara berkembang seperti Indonesia. Kondisi jalan yang baik mendukung mobilitas, memperlancar logistik, dan meningkatkan kualitas hidup masyarakat. Di Lampung, khususnya Kabupaten Lampung Tengah, kondisi jalan rusak yang menghubungkan Kabupaten Kotagajah dan Kabupaten Gayabaru telah menjadi keluhan utama masyarakat selama hampir satu dekade. Jalan berlubangdan rusak ini tidak hanya menghambat aktivitas harian tetapi juga membahayak an keselamatan pengguna jalan. Oleh karena itu, penelitian ini penting untuk memahami dampak sosial dari perbaikan jalan dan menggali opini publik yang dapat menjadi dasar kebijakan yang lebih efektif. Kondisi jalan rusak di Lampung Tengah menunjukkan kontradiksi antara harapan dan realitas. Meski telah lama diabaikan, jalan-jalan tersebut kini mendapat perhatian setelah hampir 10 tahun dibiarkan tanpa perbaikan signifikan. Perbaikan yang dilakukan oleh Kementeri an Pekerjaan Umum dan Perumahan Rakyat (PUPR) baru-baru ini mengacu pada Instruksi Presiden Nomor 3 Tahun 2023 tentang Percepatan Konektivitas Jalan Perdesaan. Meski pemer intah pusat telah mengalokasikan dana Rp 800 miliar untuk memperbaiki 15 jalan rusak di Provi nsi Lampung, masyarakat tetap resah dan skeptis mengenai efektivitas langkah ini. Perbedaan antara kondisi jalan saat ini dan kondisi ideal yang diharapkan menciptakan masalah sosial dan ekonomi yang signifikan bagi penduduksetempat [1].

Platform media sosial telah menjadi salah satu sarana utama masyarakat untuk menyampaikan pendapat dan mendiskusikan berbagai isu sosial dan politik. Komentar yang diposting di *YouTube*, salah satu platform media sosial yang paling banyak digunakan, dapat memberikan wawasan berharga mengenai opini dan sentimen publik mengenai topik tertentu. Analisis komentar YouTube terkait permintaan jalan rusak di Lampung menjadi contoh bagaimana big data dari media sosial dapat diolah untuk lebih memahami situasi sosial suatu masyarakat. Pengambilalihan jalan rusak di Lampung merupakan isu penting yang mempengar uhi kehidupan sehari - hari penduduk setempat, logistik, dan perekonomian lokal. Orang-orang menggunakan platform seperti YouTube untuk berbagi pengalaman, kekhawatiran, dan harapan mereka mengenai kondisi infrastruktur ini. Dengan menganalisis komentar-komentar yang dikumpulkan, para pemangku kepentingan seperti pemerintah daerah, pembuat kebijakan, dan organisasi masyarakat dapat mengidentifikasi tren sentimen publik, prioritas masyarakat, dan bi dang - bidang yang memerlukan perhatian dan perbaikan. *YouTube* merupakan salah satu bentuk pelaporan video dengan pengguna aktif terbanyak, pengguna dapat berkomunikasi menggunakan berbagai video, berbagi ketidaksukaan atau suka, menambah pemirsa pada video, dan berlangganan saluran [2].

Komentar pengguna dalam video dapat dianalisis untuk melihat anggapan individu terh adap kebijakan yang dilaporkan dan digunakansebagai bahan pertimbangan pembuat kebijakan. Komentar tersebut berbentuk teks, sehingga harus dilakukan analisis *text mining*. *Text mining* adalah ilmu di bidang Data Mining, yang mempelajari pemrosesan otomatis data teks dengan tujuan mengekstraksi informasi baru dari kumpulan data yang besar. Penambangan teks memberikan solusi untuk masalah seperti pengelompokan dan analisis teks tidak terstruktur dalam jumlah besar. Namun, memproses komentar untuk mengekstrak informasi yang bermakna sangat sulit, setidaknya karena dua alasan: (i) penggunaan kata dan ejaan yang tidak standar dan (ii) masalah konversi kode [3]. Teknik yang umum digunakan dalam penelitian *text mining* adalah clustering. *Clustering* merupakan suatu teknik yang digunakan untuk mengelompokkan data ke dalam suatu cluster dengan menggunakan parameter parameter tertentu sedemikian rupa sehingga objek objek dalam cluster tersebut mempunyai d erajat kemiripan yang sama [4].

Algoritma *clustering* digunakan untuk mengelompokkan dan menganalisis sentimen komentar YouTube terkait permintaan jalan rusak di Lampung. Metode siku digunakan untuk menentukan jumlah *cluster* yang optimal untuk mengelompokkan data, memungkinkan komentar diorganisasikan ke dalam kelompok-kelompok dengan karakteristik serupa dalam hal sentimen dan topik diskusi. Selanjutnya kami menerapkan algoritma *K-Means* untuk melakukan *clustering*. Hal ini menambah struktur pada data tidak terstruktur Anda dan memungkinkan analisis lebih lanjut mengenai jenis sentimen dan tema umum dalam diskusi publik tentang topik tersebut. Melalui pendekatan analitis tersebut, penelitian ini tidak hanya memberikan gambaran *holistik* mengenai sentimen dan opini masyarakat mengenai pengambilalihan jalan rusak di Lampung, namun juga mengidentifikasi *area-area* dimana para pemangku kepentingan dapat melakukan intervensi secara strategis untuk menanggapi pandangan masyarakat. Oleh karena itu, hasil analisis ini diharapkan dapat memberikan kontribusi pada proses pengambilan kebijakan yang lebih tepat guna menjawab kebutuhan dan harapan masyarakat lokal. Penelitian ini menggunak an metode *K-Means*. Metode *K-Means* merupakan salah satu teknik *clustering* data yang membagi data menjadi beberapa kelompok berdasarkan karakteristiknya [5]. Namun *algoritma K-Means* juga mempunyai kekurangan yaitu hasil *clustering* bergantung pada nilai c atau jumlah *cluster* [6]. Oleh karena itu, peneliti pada penelitian ini menggunakan metode *Elbow* untuk menentukan nilai jumlah *cluster* (c) terbaik agar tercipta *cluster* yang optimal. Metode *Elbow* merupakan metode yang umum digunakan untuk menentukan jumlah *cluster* terbaik dalam suatu kumpulan data untuk digunakan dalam proses *clustering*. Jika terdapat nilai yang menunjukkan sudut pada grafik atau jika nilai tersebut menunjukkan penurunan paling signifikan, berarti jumlah *cluster* merupakan jumlah *cluster* yang terbaik [7]. Menurut Suyanto pada tahun 2020, DBSCAN mampu menghasilkan banyak *cluster* yang bebas dan acak (tidak melingkar) serta dapat membuat *cluster* dengan lebih mudah jika terdapat *noise* pada beberapa *cluster* tersebut. Algoritma DBSCAN dapat menemukan setiap *cluster* dalam bentuk apapun dan secara efektif mengidentifikasi titik-titik *noise* yang ada [8]. Pengelompokan Spasial Hierarki Aplikasi dengan Kebisingan Berbasis Kepadatan HDBSCAN adalah algoritma analisis klaster yang banyak digu nakan karena ketahanannya terhadap kebisingan dalam kumpulan data [9].

Penelitian ini bertujuan untuk menganalisis sentimen dan opini publik mengenai

perbaikan jalan rusak di Lampung melalui komentar - komentar *YouTube*. Dengan mengelompo kkan komentar menggunakan metode clustering, penelitian ini diharapkan dapat memberikan gambaran holistik tentang sentimen masyarakat dan mengidentifikasi bidang-bidang yang memerlukan perhatian lebih lanjut. Manfaat dari penelitian ini adalah menyediakan data yang dapat digunakan oleh pemerintah daerah, pembuat kebijakan, dan organisasi masyarakat sipil untuk membuat keputusan yang lebih tepat dan responsif terhadap kebutuhan masyarakat. Dengan demikian, penelitian ini berkontribusi pada proses pengambilan kebijakan yang lebih efektif dan efisien, sekaligus meningkatkan transparansi dan akuntabilitas dalam pelaksanaan perbaikan jalan.

2. Tinjauan Pustaka

Pada penelitian pertama menganalisis sentimen masyarakat Indonesia terhadap eksistensi K-Pop pada media sosial *twitter.* Kinerja dari sebuah algoritma klasifikasi dipengaruhi dari jenis data dan fiturnya, maka dari itu data set berupa teks yang akan diolah harus melalui tahapan *Text preprocessing* seperti *case folding, stemming, tokenizing, Text Normalization* serta *stopwords*, lalu setelah itu data akan masuk tahapan selanjutnya yaitu tahapan klasifikasi menggunakan algoritma *K-Means* dan diuji dengan perhitungan *Silhouette Coefficient* untuk mendapatkan nilai akurasi yang sesuai dengan harapan sehingga dapat mengklasifikasikan data untuk mendapatkan hasil kesimpulan [10].

Penelitain kedua dengan metode yang sama yaitu Implementasi dari pendekatan metode K-Means untuk mengetahui kecenderungan opini masyarakat terhadap pemilu dalam media sosial twitter. Pada penelitian ini dilakukan untuk mengetahui kecenderungan opini masyarakat terhadap pemilu apakah termasuk kedalam sentimen positif atau *negative* [11].

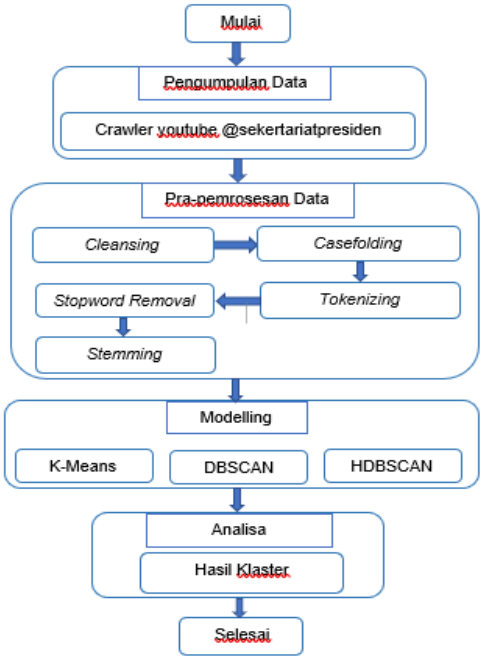
Penelitain ketiga melakukan penelitian pada tahun 2020 tentang “Analisis Klasifikasi Sentimen Pengguna Media Sosial *Twitter* Terhadap Pengadaan Vaksin COVID-19”. Penelitian ini bertujuan untuk mengetahui bagaimana sentimen publik terhadap sebuah masalah atau objek, apakah cenderung beropini negatif atau positif. Penelitian ini melakukan pengumpulan data dengan melakukan *crawling twitter* dan menghasilkan 1000 tweet untuk dataset. Dan menghasilkan persentase opini masyarakat terhadap vaksin Corona yaitu 48% positif, 29% netral, dan 23% negative [12].

Penelitian keempat mengenai Analisis Sentimen Publik pada media social Twitter mengenai pelaksanaan pilkada serentak menggunakan algoritma *K-Means* dan *Support Vector Machine*. Penelitian ini bertujuan untuk mengetahui respon masyarakat pada media sosial Twitter tentang kelangsungan pilkada. Penelitian ini menggunakan 3000 *tweet* Bahasa Indonesia yang digunakan untuk dataset dan membagi data kedalam 2 kategori yaitu *Cluster* 1 sebagai kelompok *Tweet* positif dan *Cluster* 2 sebagai kelompok *Tweet* negatif [13].

Pada penelitian ini akan dilakukan *clustering* review pengguna aplikasi Zenius pada layanan Google Play Store menggunakan metode K-Means, DBSCAN (Density-Based Spatial Clustering of Application with Noise) dan metode HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) dan membadingkan ketiga metode tersebut menggunakan *silhouette coefficient*. Adapun tujuan dari penelitian ini adalah untuk mengisi gap tersebut dengan membandingkan performa K-Means, DBSCAN, dan HDBSCAN dalam mengelompokkan analisis sentimen komentar YouTube terkait pengambilalihan jalan rusak di Lampung. Dengan melakukan perbandingan ini, diharapkan dapat diidentifikasi algoritma yang paling efektif dan efisien dalam menangani data sentimen dengan karakteristik yang beragam.

3. Metodologi

Perencanaan diagram analisis sentimen adalah gambaran yang menunjukkan alur penelitian yang dilakukan. Alur pencarian bisa dilihat pada gambar 1.



Gambar 1. Tahapan Metode Penelitian

3.1. Pengumpulan Data

Data komentar berasal dari akun *YouTube* @sekretariatpresresiden. Website ini membagikan konten hasil kunjungan presiden ke berbagai daerah di Indonesia dan menggunakan izin layanan *Youtube* Data API v3 untuk mengakses atribut data statistik video, seperti komentar nama orang, konten komentar, tanggal komentar, jumlah like, dan jumlah balasan. Jumlah data komentar dari seluruh link yang diperoleh cukup besar, yaitu 10814 baris data tidak berlabel. Pengumpulan data dilakukan dengan menggunakan bahasa pemrograman Python yang dimodifikasi menggunakan perpustakaan *Selenium Webdriver*.

3.2. Pra-pemrosesan Data

Pemrosesan awal teks merupakan langkah penting dalam penambangan teks. Langkah *preprocessing* memproses data mentah menjadi data bersih untuk menyederhanakan proses *clustering*. Langkah-langkah prapemrosesan dokumen meliputi:

3.2.1. Cleansing

Pengaplikasian pembersihan dataset yang dilakukan bertujuan untuk menghilangkan tanda baca yang tidak diperlukan sehingga dapat mengekstrak pola-pola yang potensial [14]**.**

3.2.2. Case Folding

Merupakan proses penyalarasan teks pada kalimat yang ada pada dataset. Pada penelitian ini, data akan dilakukan Case Folding pada dataset adalah lowercase, yaitu dengan mengubah semua data teks yang telah didapat menjadi huruf kecil [14].

3.2.3. Tokenizing

Merupakan proses pemisahan antara teks dalam kalimat berdasarkan spasi atau pun symbol [14], seperti kalimat “pemimpin disana ngapain aja smpe mesti pusat turun kesana parah sih” menjadi list ['pemimpin', 'disana', 'ngapain', 'aja', 'smpe', 'mesti', 'pusat', 'turun', 'kesana', 'parah', 'sih'] yang awalnya berbentuk satu kalimat utuh menjadi per kata.

3.2.4. Stopword Removal

Menghapus *stopword* akan menghilangkan kata-kata yang dianggap tidak penting didalam teks [15], seperti imbuhan dan pronoun seperti “it” dan “they”.

3.2.5. Stemming

Stemming akan mengubah kata yang ada pada komentar menjadi bentuk dasarnya sehingga dapat mengurangi variasi fitur-fitur yang memiliki makna yang sama namun karena terdapat imbuhan pada kata tersebut menyebabkan adanya perbedaan makna kata [16].

3.3. Pembobotan Kata

Setelah tahap preprocessing teks membuat sekumpulan istilah atau kata, langkah selanjutnya adalah pembobotan kata, dimana setiap kata akan diberi bobot atau nilai. Bobot atau nilai akan menunjukkan pentingnya sebuah kata dalam komentar. Tujuannya untuk mengetahui persamaan dan ketersediaan suatu kata pada komentar. Semakin sering sebuah kata muncul, semakin besar bobot atau nilainya. Dalam proses penghitungan bobot kata, metode yang digunakan adalah metode TF-IDF. *Term Frekuensi Invers Frekuensi* (TF-IDF) adalah metode algoritmik yang berguna untuk menghitung bobot atau nilai setiap kata yang umum digunakan. TF-IDF mengevaluasi pentingnya sebuah kata dalam sebuah dokumen. Hal ini tergantung pada berapa kali kata tersebut muncul dalam dokumen [17]. Persamaan pembentukan TF-IDF dapat dilihat pada persamaan 1 dan 2 di bawah ini.

$$W_{i,j} = TF_{i,j}\ x\ IDF_j \quad \ldots\ldots\ldots(1)$$

$$\mathrm{ID}_j = \log\left(\frac{N}{DF_j}\right) \quad \ldots\ldots\ldots(2)$$

Keterangan:
$W_{i,j}$= bobot dari kata ke j pada komentar ke i
ID_j = banyaknya komentar yang mengandung kata j
$TF_{i,j}$ = jumlah kemunculan kata ke j pada komentar ke i
IDF_j= inverse document frequency pada kata ke j
N = jumlah keseluruhan komentar

2.4. Modelling

Pemodelan merupakan tahap pemilihan model dan penerapan pemodelan dengan algoritma *data mining*. Tahap ini bertujuan untuk mengoptimalkan hasil penelitian. Pada penelitian ini terdapat 3 pemodelan yang dilakukan.

2.4.1. Elbow

Metode *Elbow* merupakan salah satu metode untuk menentukan jumlah *cluster* yang tepat melalui persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik [18]. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka jumlah nilai *cluster* tersebut yang tepat. Untuk mendaatkan perbandingannya adalah dengan menghitung *Sum of Square Error* (SSE) dari masing-masing nilai cluster. Karena semakin besar jumlah nilai cluster K, maka nilai SSE akan semakin kecil. Rumus SSE sesuai dengan Persamaan 3.

$$SC = \sum_{k=1}^{n} \sum_{x_i} |x_{i\ -}\ c_{k|2} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

Keterangan:
K = cluster ke-c
xi = jarak data obyek ke-i
ck = pusat cluster ke-i

2.4.2. K-Means

Algoritma *K-Means* merupakan teknik analisis data yang menggunakan sistem partisi untuk melakukan proses pengelompokan data [17]. Suatu data dikelompokkan ke dalam satu *cluster* berdasarkan kemiripan atribut yang dimiliki. Kemiripan ini bisa diketahui dengan mengukur jarak setiap data dengan pusat *cluster* (*centroid*) [5]. Metode *K-Means Clustering* merupakan salah satu metode *Clustering* untuk mengelompokkan data yang memiliki jumlah data besar dan

proses yang cepat dan efisien [19]. Berikut adalah langkah perhitungan algoritma *K-Means:*
1).Tentukan berapa banyak jumlah cluster(c) atau kelompok.
2).Tentukan secara acak pusat cluster awal (centroid).
3).Ukur jarak setiap data dengan pusat cluster (centroid). yaitu dengan rumus Euclidean Distance pada persamaan 4.

$$D_{i,j} = \sqrt{(x_{1p} - x_{pq})^2 + (x_{2p} - x_{2q})^2} + +(x_{rp} - x_{rq})^2 \quad \text{..... (4)}$$

Keterangan:
$D_{(p,q)}$ = jarak data ke-p dengan pusat cluster q
$X_{(r,p)}$ = data ke-p pada atribut data ke-r
$X_{(r,q)}$ = titik pusat ke-q pada atribut

4. Kelompokkan setiap data ke dalam cluster berdasarkan jarak minimum
5. Lakukan proses iterasi dengan menentukan pusat klaser (centroid) baru dengan rumus pada persamaan 5.

$$SC = \sum_{k=1}^{n} \sum_{x_i} |x_i - c_{k|2} \quad \text{..... (5)}$$

Keterangan:
v = centroid pada cluster
$x(p)$= objek ke-p
n= banyaknya objek/jumlah

6). Ulangi langkah 3 hingga tidak terdapat perubahan *cluster* pada setiap data dari proses iterasi sebelumnya.

2.4.3. DBSCAN

Algoritma DBSCAN dapat menemukan sampel inti dengan kepadatan tinggi dan memperluas *cluster* dari sampel tersebut. Terdapat dua parameter utama dari algoritma yang menentukan *cluster*: jumlah sampel minimal dan ε. Parameter pertama menentukan jumlah titik minimal yang dapat diklasifikasikan bersama sebagai sampel inti. Parameter ini mendefinisikan tingkat toleransi noise dari algoritma [20].
1) Tentukan nilai parameter MinPts dan Eps.
2) Tentukan secara acak nilai p atau titik awal.
3) Hitung Eps atau semua jarak titik yang density reachable terhadap p menggunakan rumus jarak euclidean berikut.

$$D_{ij} = \sqrt{\sum_{a}^{p} (x_{ia} - x_{ja})^2} \quad \text{..... (6)}$$

Dimana x_{ia} merupakan variabel ke-a dari obyek i (i=1, …, n; a=1, …, p) dan D_{ij} adalah nilai euclidean distance.
4) Terbentuk sebuah cluster ketika titik yang memenuhi Eps lebih dari MinPts dan titik p sebagai core point.
5) Lakukan pengulangan langkah 3 – 4 hingga dilakukan proses pada semua titik. Jika p merupakan titik border dan tidak ada titik yang density reachable terhadap p, maka proses dilanjutkan ke titik yang lain.

2.4.4. HDBSCAN

Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) a dalah algoritma analisis klaster yang banyak digunakan karena ketahanannya terhadap noise d alam kumpulan data [9].

2.4.5. Silhoutte Coefficient

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik atau buruknya suatu obyek ditempatkan dalam suatu *cluster*. Metode ini merupakan

gabungan dari metode separasi dan kohesi [21]. Untuk menghitung nilai *silhouette coefficient*, diperlukan perhitungan nilai *silhouette index* dari sebuah data ke-i. Nilai *silhouette coefficient* didapatkan dengan mencari nilai maksimal dari nilai *Silhouette Index Global* dari jumlah *cluster* 2 sampai jumlah *cluster* n-1, seperti pada Persamaan 7 berikut.

$$SC = maks_k\ SI\ (k) \qquad \text{……..(7)}$$

Untuk menghitung nilai SI dari sebuah data ke-i, ada 2 komponen yaitu ai dan bi. Nilai ai adalah rata-rata jarak ke-i terhadap semua data lainnya dalam satu cluster, sedangkan bi didapatkan dengan menghitung rata-rata jarak data ke-i terhadap semua data dari *cluster* lainnya yang tidak satu *cluster* dengan data ke-i, lalu diambil yang terkecil [22]. Berikut Persamaan 5 untuk menghitung nilai $ai\ j$.

$$a_i^j = \frac{1}{mj-1} \sum_{\substack{r=1 \\ r\neq 1}}^{m_j} d(x_i^j, x_r^j) \qquad \text{……(8)}$$

Keterangan:
j = *cluster* i = index data (i = 1,2,…mj) ai
j = rata-rata jarak data ke-i terhadap semua data dalam satu cluster Mj = jumlah data dalam *cluster* ke-j
$d(xi\ j\ ,\ xr\ j\ )$ = jarak data ke-i dengan data ke-r dalam satu cluster j.
Berikut ini adalah rumus perhitungan mendapatkan nilai $bi\ j$ dapat dilihat pada Persamaan 9.

$$b_i^j = min_{\substack{n=1,..k \\ n\neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r\neq 1}}^{m_n} d(x_i^j, x_r^n) \right\} \qquad \text{…….. (9)}$$

Keterangan:
j = cluster
i = indexdata(i=1,2,…mj)
b_i^j = rata-rata jarak data ke-i terhadap semua data
yang tidak dalam satu *cluster* dengan data ke-i
m_n= jumlah data dalam cluster ke-n
$d(x_i^j,\ x_r^n)$= jarak data ke-i dengan data ke-j dalam satu *cluster* n
Berikut ini adalah rumus perhitungan mendapatkan nilai SI_i^j dapat dilihat pada Persamaan 10.

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \qquad \text{…… (10)}$$

Keterangan:
SI_i^j = *Silhoutte Index* data ke-i dalam satu *cluster*
b_i^j = rata-rata jarak data ke-i terhadap semua data yang tidak dalam satu cluster dengan data ke-i
a_i^j = rata-rata jarak data ke-i terhadap semua data dalam satu cluster.
Berikut ini adalah rumus perhitungan mendapatkan nilai SI_j dapat dilihat pada Persamaan 11

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \qquad \text{…… (11)}$$

Keterangan:
SI_j = Rata-rata *Sillhouette Index cluste*r j
SI_i^j= *Silhoutte Index* data ke-i dalam satu *cluster*
M_j = jumlah data dalam cluster ke-j
i = index data (i = 1,2,… m_j)
Berikut ini adalah rumus perhitungan mendapatkan nilai SI global sesuai dengan Persamaan 12.

$SI = \frac{1}{K} \sum_{j=1}^{k} SI_j$ ..... (12)

Keterangan:
SI = Rata-rata *Sillhouette Index* dari dataset
SIj = Rata-rata *Sillhouette Index cluster* j
k = jumlah *cluster*

4. Hasil dan Pembahasan

Data yang berhasil dikumpulkan dari akun youtube @serketariatpresiden adalah seban yak 10814 komentar, data tersebut dikumpulkan dari satu video youtube yang berkaitan dengan Kunjungan Presiden Jokowi keLampung Tengah. Gambar 2 menunjukkan dataset komentar youtube yang digunakan.

	tanggal	nama	komentar	suka
0	2024-03-22T01:53:43Z	@user-fb6vk2ut8d	Pak presiden tolong lh jalan kami di des...	0
1	2024-03-22T01:22:51Z	@Yokoubus	Arus jalan itu jagan kuwat makan uwan	0
2	2024-03-22T01:21:50Z	@Yokoubus	Adu jalan tida bagus	0
3	2024-03-15T22:03:05Z	@adieprayoga7186	GUBERNUR lampung malah tepuk tangan, saya sbag...	0
4	2024-03-15T04:13:51Z	@user-tb7gi3ip7k	Gubernur lampung gatau diri perut besar ketawa...	0
...	...	...	...	...
9995	2023-05-05T15:18:47Z	@beng6231	Biar bagemanapun itu kan gubernur pilihan kali...	0
9996	2023-05-05T15:18:46Z	@gatotrudijanto5721	Siapa yang pilih .....yooo?	0
9997	2023-05-05T15:18:46Z	@MASDANANGSUPRAYITNO	KPK HARUS TURUN TANGAN JUGA ,PECAT PENJABAT NA...	0
9998	2023-05-05T15:18:45Z	@angusyoung8927	Itu gubernur masih bisa cengar cengir...ga mal...	0
9999	2023-05-05T15:18:44Z	@ariwimantono4829	<a href="https://www.youtube.com/watch?v=Wvg2H...	0

Gambar 2. Hasil Crawling Data

3.2. Preprosesing Data

Prapemrosesan teks atau yang lebih dikenal dengan nama *text preprocessing* yaitu pro ses membersihkandata sebelum diolah nantinya. Pada tahapan ini terdapat 5 proses diantaranya.

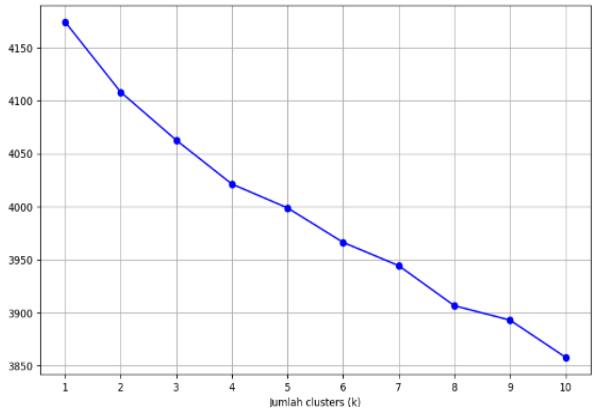
Tabel 1. Hasil Processing Data

Preprosesing	Input	Output
Cleaning	Sampe2 presiden turun langsung. Pejabat Disindir tapi malah cengar cengir ada ada aja.	Sampe presiden turunlangsungbr Pejabat disindir tapi malah cengar cengir ada ada aja.
CaseFolding	Sampe presiden turun langsung b r pejabatengapain aja sihdimana h arus presiden yg pantau lgsg ckck ck	disindir tapi malah cengar cengir ada ada aja
Tokenizing	Sampe presiden turun langsung b r pejabatengapain aja sihdimana h arus presiden yg pantau lgsg ckck ck	['sampe','presiden','turun','langsun gbrpejabat','e','ngapain','aja','sih',' dimana','harus','presiden','yg','pan tau','lgsg','ckckck']

Preprosesing	Input	Output
Stopword Removal	['gubernur', 'muka','badaktau', 'malu','gak','sih','dia','stlh','di', 'sindir', 'sama','presiden', 'langsung','dan', 'menteri', 'moga','keluarga','nya','tidak','punya','malubrbrkomuknya','banyak','makan','duit','haram','jd','jabat','bukan','krn','ingin','baik','daerah','tp','krn','ingin','kaya','diri','sendiri']	gubernur muka badaktau malu gak sih stlh sindir presiden langsung menteri moga keluarganya malubrbrkomuknya makan duit haram jd jabat krn daerah tp krn kaya
Stemming	['saya','orang','lampung','lahir','disana','alhamdulillah','kalo','dibngaun','jalannya','gubernur','urat','malunya','udah', 'putuss']	['saya','orang', 'lampung','lahir', 'sana','alhamdulillah','kalo','dibngaun','jalan','gubernur','urat','malu','udah','putuss']

3.2. Modelling
3.2.1. Metode K-means

Setelah data siap dipakai langkah selanjutnya yaitu melakukan pemodelan. Pada penelitian ini dalam menentukan jumlah *cluster* (c) yang optimal adalah menggunakan metode *elbow*. Dengan metode *elbow* jumlah cluster dikatakan optimal apabila grafik perbandingan *Sum of Square Error* (SSE) dengan jumlah *cluster* membentuk siku (*elbow*). Artinya, selisih nilai SSE dengan cluster sebelumnya berubah drastis namun selisih nilai SSE dengan *cluster* setelahnya tidak berubah drastis. Di bawah ini tabel 2 merupakan nilai SSE dengan jumlah *cluster* mulai dari c=1 sampai c=10.



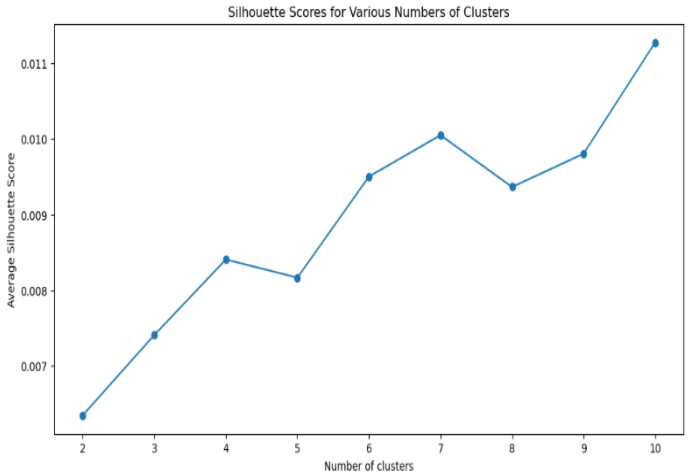
Gambar 4. Hasil Grafik Elbow

Berdasarkan gambar hasil pengujian menggunakan *elbow method* maka jumlah *cluster* yang baik yang digunakan adalah 3 *cluster*, sehingga dalam penelitian ini menggunaka 3 *cluster* yaitu cluster 0, *cluster* 1, dan *cluster* 2.

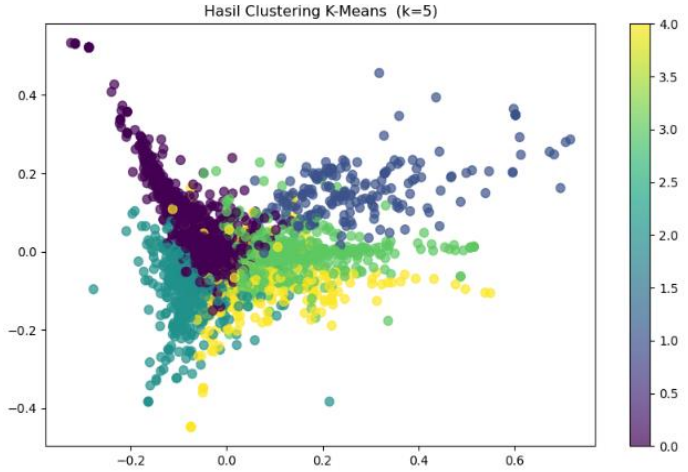
Tabel 2. Tabel Hasil Silhoutte Coefficient

No	Jumlah Cluster (c)	Skor Silhouette
1	n_clusters= 2 Skor Silhouette rata-rata	0.01554175776108209
2	n_clusters= 3 Skor Silhouette rata-rata	0.0167349555981015
3	n_clusters= 4 Skor Silhouette rata-rata	0.018883212038088012
4	n_clusters= 5 Skor Silhouette rata-rata	0.018753402264614038
5	n_clusters= 6 Skor Silhouette rata-rata	0.0206880183084802

Berdasarkan gambar hasil pengujian menggunakan *silhouette coefficient* maka jumlah cluster yang baik yang digunakan adalah 3 *cluster* dengan nilai *silhouette coefficient* yaitu 0.187, dan lebih baik dari nilai *silhouette coefficient cluster* lainnya.



Gambar 4. Hasil Grafik *Silhoutte Coefficient*



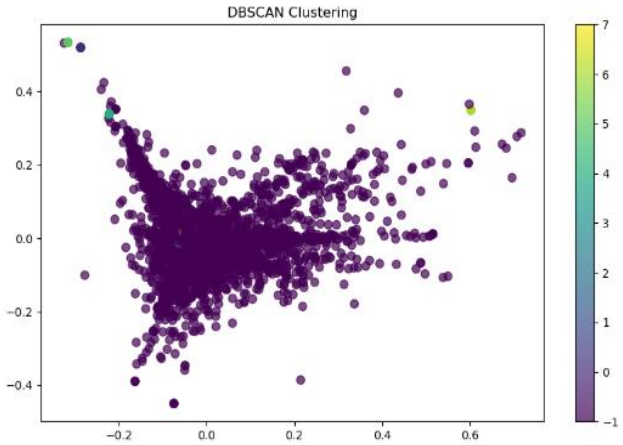
Gambar 5. Plot Hasil *Clustering K-means*

3.2.2. DBSCAN

	text	cluster
0	ajar gubernur lampung yg full senyum	-1
1	bukan viral viralnya tugas i love presiden jokowi	-1
2	hahaha kasi jokowi	-1
3	minimal jalan kabupaten kota camat bagus wajib...	-1
4	presiden best is the best	7
...	...	...
995	gak benah jalan nya ikan lele nya mati	-1
996	sngat pangkat nya provinsi la walikota la bupa...	-1
997	baju coklat muka senyum tepuk tangan untung sa...	-1
998	lampung jawa brlogat nya loh	-1
999	presiden jokowi jalan bopreng	-1

1000 rows × 2 columns

Gambar 6. Hasil *Clustering* DBSCAN



Gambar 7. Plot Hasil *Clustering* DBSCAN

3.2.3. HDBSCAN

```
Jumlah kluster yang ditemukan: 110
Skor Silhouette: 0.10672286827790585

Clustering:
Teks dari kluster pertama:
denger milliar neng banget mr gov gin nih kalo otonomi daerah bablas resin perintah pusat hihihi

Teks dari kluster kedua:
gatau malu gubernur ga becus kerja cengengesan be ahun jalan rusak parah sih orang

Teks dari kluster ketiga:
lihat sragam coklat d smping presiden cari muka nyadisindir mlh tepuk tngan lo ha lo hopintar pintar
berdramabrmoga aja kpk periksa intansi

Teks dari kluster keempat:
gin rakyat indonesia njir jokowi sindir jalan nya hancur udah parah tidur nyenyak mobil asli ya

Teks dari kluster kelima:
hehehe ngibul
```

Gambar 8. Hasil Clustering HDBSCAN

3.2. Hasil Analisa

Nilai skor siluet K-means sebesar -0,3484. Nilai negatif menunjukkan bahwa *cluster* yang dihasilkan K-Means kurang serupa atau datanya mungkin tidak sesuai dengan model *cluster* yang digunakan. Skor siluet DBSCAN adalah 0,8368. Nilai yang tinggi menunjukkan bahwa DBSCAN mampu menghasilkan *cluster* dengan kemiripan antar anggota yang tinggi, dan *cluster* tersebut mempunyai batasan yang jelas dibandingkan dengan *cluster* lainnya. Nilai skor siluet HDBSCAN sebesar 0,1067. Meski lebih rendah dibandingkan DBSCAN, nilai positif ini menunjukkan bahwa HDBSCAN meski tidak sekuat DBSCAN, namun masih mampu menemukan struktur *clustering* yang sangat baik pada data.

Tabel 3. Tabel Hasil Skor *Silhouette* Dari 3 Algoritma

K-Means	DBSCAN	HDBSCAN
-0.34845129089199744	0.8368084953566132	0.10672286827790585

3.3. Pembahasan

Hasil pengujian menunjukkan bahwa DBSCAN adalah algoritma *clustering* yang paling efektif untuk dataset ini, dengan skor tertinggi dan kemampuan menangani noise serta variasi

densitas yang baik. Hasil ini sejalan dengan penelitian terdahulu yang menegaskan keunggulan DBSCAN dalam berbagai aplikasi data spasial dan data kompleks. K-Means terbukti tidak efektif, konsisten dengan banyak penelitian sebelumnya yang menunjukkan keterbatasannya pada dataset *non-sferis* dan dengan *noise*. HDBSCAN memberikan hasil yang baik tetapi masih kurang optimal dibandingkan DBSCAN, menunjukkan bahwa parameter dan kondisi dataset sangat mempengaruhi performa algoritma *clustering* ini.

5. Simpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa dataset berupa kumpulan komentar dari youtube tentang pengelompokan komentar *YouTube* mengenai penga mbilalihan jalan rusak diLampun sebanyak 4353 komentar. Nilai *Silhouette Score* tertinggi diper oleh dengan menggunakan algoritma DBSCAN sebesar nilai akurasi 0,836%.

Daftar Referensi

[1] A. N. H. Regita and I. Santoso, "Analisis Sentimen Publik Terhadap Pengambilalihan Jalan Rusak Di Lampung Menggunakan Algoritma K-Nearest Neighbors (KNN)," IKRA-ITH Informatika:Jurnal Komputer dan Informatika, vol. 7, no. 2, pp. 176-182, 2023.

[2] F. Abdulloh, F. Ferian, and I. R. Pambudi, "Analisis sentimen pengguna YouTube terhadap program vaksin COVID-19," Csrid (Computer Science Research and Its Development Journal), vol.13,no.3,pp. 141-148, 2021.

[3] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of Indonesian-English code-mixed Twitter data," in Proc. 5th Workshop on Noisy User-generated Text (W-N T), 2019, pp. 1-10.

[4] M. Z. Fauzi and A. Abdullah, "Clustering of public opinion on natural disasters in Indonesia using DBSCAN and K-Medoids algorithms," Journal of Physics: Conference Series, vol. 1783, no.1,pp.1-7, 2021.

[5] A. Ali, "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo," MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 19, no. 1, pp. 186-195, 2019.

[6] A. P. Riani, A. Voutama, and T. Ridwan, "Penerapan K-Means clustering dalam pengelompokan hasil belajar peserta didik dengan metode Elbow," Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD, vol. 6, no.1,pp.164-172,2023.

[7] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," Matrix: Jurnal Manajemen Teknologi dan Informatika,vol.9,no.3,pp.102-109,2019.

[8] W. Jing, C. Zhao, and C. Jiang, "An improvement method of DBSCAN algorithm on cloud computing," Procedia Computer Science,vol.147, pp. 596-604, 2019.

[9] I. Ghamarian and E. A. Marquis, "Hierarchical density-based cluster analysis framework for atom probetomography data,"Ultramicroscopy, vol. 200, pp. 28-38, 2019.

[10] I. K. A. Wirayasa and H. Santoso, "Analisis Employee Satisfaction Menggunakan Teknik Clustering Dan Classification Machine Learning," Progresif: Jurnal Ilmiah Komputer, vol. 18, no.1,pp.1-10,2022.

[11] I. Kurniawan, I. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," Jurnal Eksplora Informatika, vol. 9, no.1,pp.1-10,2019.

[12] M. I. Aditama, et al., "Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19," *Journal Information Engineering and Educational Technology*, vol. 869X,no.2549,2020..

[13] Rahmawati, Atika, Aris Marjuni, and Junta Zeniarja. "Analisis Sentimen Publik Pada Media Sosial Twitter Terhadap Pelaksanaan Pilkada Serentak Menggunakan Algoritma Support Vector Machine." Creative Communication and Innovative Technology Journal 10.2 (2017): 197-206.

[14] A. G. Arja, "Penerapan Sentimen Analis Menggunakan Metode Naïve Bayes Dan SVM," Jurnal Ilmu Data,vol. 2, no. 12, 2022.

[15] S. Khairunnisa, A. Adiwijaya, and S. A. Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),"Jurnal Media Informatika Budidarma, vol. 5, no. 2, pp. 406-414, 2021.

[16] I. M. A. Purniawan, G. M. A. Sasmita, and I. P. A. E. Pratama, "Clustering Berita

Menggunakan Algoritma Tf-Idf dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.com," Jurnal Ilmiah Teknologi dan Komputer, vol. 3, no. 1, pp. 821-830, 2022.

[17] M. Z. Fauzi and A. Abdullah, "Clustering of public opinion on natural disasters in Indonesia using DBSCAN and K-Medoids algorithms," Journal of Physics: Conference Series, vol. 1783,no.1,pp.1-7, 2021.

[18] W. Saefudin, A. Komarudin, and R. Ilyas, "Visualisasi Kumpulan Berita Dalam Bentuk Peta Digital Dengan Metode Term Frequency-Inverse Document Frequency dan Gazetteer," in Proc.Seminar Nasional Sains dan Teknologi Informasi (SENSASI), vol. 2, no. 1, 2019.

[19] M. Z. Fauzi and A. Abdullah, "Clustering of public opinion on natural disasters in Indonesia using DBSCAN and K-Medoids algorithms," Journal of Physics: Conference Series, vol. 1783,no.1,pp.1-7, 2021.

[20] N. P. E. Merliana and A. J. Santoso, "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means Clustering," presented at the Conf., 2015.

[21] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," Matrix:Jurnal Manajemen Teknologi dan Informatika, vol. 9, no. 3, pp. 102-109, 2019.

[22] T. Purwanti, W. Ramdhan, and S. Santoso, "Penerapan Metode Klasterisasi K-Means untuk Strategi Promosi Pada SMK Tamansiswa Sukadamai," JUTSI: Jurnal Teknologi dan Sistem Informasi,vol.1,no. 2, pp. 141-146, 2021.