TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi
Vol. 6 No. 1 (Juni 2026)
https://doi.org/10.46880/tamika.Vol6No1.pp20-29 e-ISSN: 2828-1276

PERBANDINGAN METODE K-MEANS DAN DBSCAN UNTUK CLUSTERING TEKS PESAN MENFESS DI TELEGRAM

[1]Oktama Pangestu✉, [1]Abdul Hamid Arribathi, [1]Nur Azizah, [2]Rahmat Hidayat
[1]Program Studi Magister Teknologi Informatika, Universitas Raharja, Tangerang, Indonesia
[2]Jurusan Ilmu Komputer, Universitas Riau, Pekanbaru, Indonesia
Email: oktama@raharja.info

ABSTRACT

Menfess messages on Telegram are a form of anonymous communication that generates large amounts of informal text data with linguistic characteristics such as slang, abbreviations, and spelling variations, posing challenges for computational text analysis. This study compares the performance of K-Means and DBSCAN in clustering menfess messages using Sentence-BERT embedding through the paraphrase-multilingual-MiniLM-L12-v2 model across three text length scenarios: short (0–20 words), medium (55–128 words), and long (129–283 words), derived via word_count clustering. Evaluation uses Silhouette Score and Davies-Bouldin Index. For short texts, K-Means achieves 0.0804 and 3.8451, while DBSCAN produces 2 clusters with 0.3186, 1.3714, and 71.60% noise. For medium texts, K-Means obtains 0.1403 and 3.6490, while DBSCAN forms 1 cluster with 0.0450, 3.3405, and 74.60% noise. For long texts, K-Means obtains 0.0593 and 3.4552, while DBSCAN produces 2 clusters with 0.5492, 1.1069, and 79.67% noise. Results show that DBSCAN outperforms on short and long texts by evaluation metrics but produces very high noise across all scenarios, while K-Means demonstrates more stable performance by clustering all data without noise in every scenario.

Keyword: Text Clustering, K-Means, DBSCAN, Telegram, Menfess, Sentence-BERT.

ABSTRAK

Pesan menfess di Telegram merupakan bentuk komunikasi anonim yang menghasilkan data teks informal dalam jumlah besar dengan karakteristik kebahasaan berupa slang, singkatan, dan variasi ejaan, sehingga menimbulkan tantangan dalam analisis teks secara komputasional. Penelitian ini membandingkan kinerja K-Means dan DBSCAN dalam clustering teks pesan menfess menggunakan Sentence-BERT embedding melalui model paraphrase-multilingual-MiniLM-L12-v2 serta menganalisis pengaruh panjang teks terhadap kualitas cluster. Dataset dibagi ke dalam tiga skenario berdasarkan panjang teks: teks pendek (0–20 kata), menengah (55–128 kata), dan panjang (129–283 kata), yang diperoleh melalui clustering pada word_count. Evaluasi menggunakan Silhouette Score dan Davies-Bouldin Index. Pada skenario pendek, K-Means memperoleh Silhouette Score 0,0804 dan Davies-Bouldin Index 3,8451, sedangkan DBSCAN menghasilkan 2 cluster dengan 0,3186, 1,3714, dan noise rate 71,60%. Pada skenario menengah, K-Means memperoleh 0,1403 dan 3,6490, sedangkan DBSCAN menghasilkan 1 cluster dengan 0,0450, 3,3405, dan noise rate 74,60%. Pada skenario panjang, K-Means memperoleh 0,0593 dan 3,4552, sedangkan DBSCAN menghasilkan 2 cluster dengan 0,5492, 1,1069, dan noise rate 79,67%. Hasil menunjukkan bahwa DBSCAN unggul pada teks pendek dan panjang berdasarkan metrik evaluasi, namun menghasilkan noise yang sangat tinggi di seluruh skenario, sedangkan K-Means menunjukkan performa yang lebih stabil dengan berhasil mengelompokkan seluruh data tanpa noise pada setiap skenario.

Kata Kunci: Text Clustering, K-Means, DBSCAN, Telegram, Menfess, Sentence-BERT.

PENDAHULUAN

Perkembangan media sosial dan platform percakapan digital telah menghasilkan data teks dalam jumlah sangat besar. Berbagai penelitian telah memanfaatkan data teks media sosial berbahasa Indonesia untuk analisis komputasional, di antaranya analisis sentimen pada platform X menggunakan Support Vector Machine (Hidayat et al., 2024), pengelompokan konten pemasaran digital (Larasati et al., 2021), dan clustering opini pengguna media sosial (Kurniawan & Achmadi, 2024). Salah satu bentuk komunikasi digital yang berkembang pesat adalah pesan menfess di Telegram, yaitu pesan anonim yang dikirim melalui bot atau akun perantara untuk dipublikasikan ke kanal atau grup tertentu. Sifat anonim pada menfess mendorong pengguna untuk menyampaikan pertanyaan, keluhan, pengalaman pribadi, maupun opini secara lebih bebas, sehingga pesan menfess menjadi sumber data teks yang kaya dan dinamis.

Meskipun demikian, pesan menfess umumnya ditulis dalam bentuk tidak terstruktur dan didominasi

oleh bahasa tidak baku, seperti slang, singkatan, serta variasi ejaan (Budiasa, 2021; Hutauruk et al., 2024). Pesan-pesan yang memiliki makna serupa sering kali diekspresikan dengan pilihan kata yang berbeda, sehingga proses analisis dan pengelompokan teks menjadi lebih kompleks apabila hanya mengandalkan kesamaan kata secara permukaan.

Text clustering merupakan teknik dalam data mining yang digunakan untuk mengelompokkan dokumen berdasarkan kemiripannya tanpa label kelas (Adawiyah, 2023). Dalam konteks menfess, clustering dapat dimanfaatkan untuk mengidentifikasi tema-tema dominan dan pola diskusi. K-Means telah diterapkan pada pengelompokan teks media sosial seperti ulasan e-commerce (Widjaja, 2023) dan konten pemasaran digital (Larasati et al., 2021; Ompo & Pakereng, 2024). Sebaliknya, DBSCAN mengelompokkan data berdasarkan kepadatan dan unggul dalam mendeteksi noise (Ester et al., 1996), serta telah digunakan untuk deteksi topik pada data media sosial (Huang et al., 2022).

Sebagian besar penelitian sebelumnya masih menggunakan TF-IDF yang memiliki keterbatasan dalam menangkap kemiripan semantik. Reimers dan Gurevych (2019) memperkenalkan Sentence-BERT sebagai pendekatan embedding yang mampu merepresentasikan kedekatan makna antarkalimat secara lebih efektif. Namun, penerapan Sentence-BERT untuk clustering teks komunikasi anonim berbahasa Indonesia, khususnya menfess, masih terbatas.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan membandingkan kinerja K-Means dan DBSCAN dalam clustering teks pesan menfess di Telegram menggunakan Sentence-BERT embedding pada tiga skenario panjang teks, yaitu pendek, menengah, dan panjang, dengan evaluasi menggunakan Silhouette Score dan Davies-Bouldin Index.

KAJIAN LITERATUR

Menfess dan Komunikasi Anonim di Telegram

Menfess merupakan singkatan dari mention confess, yaitu pesan anonim yang dikirimkan pengguna melalui bot perantara untuk kemudian dipublikasikan di kanal atau grup Telegram. Fenomena ini berkembang pesat di kalangan pengguna berbahasa Indonesia sebagai media untuk menyampaikan perasaan, pengalaman, maupun opini tanpa mengungkapkan identitas. Sifat anonim pada menfess mendorong pengguna menulis secara lebih bebas dengan bahasa sehari-hari yang kaya akan slang, singkatan, dan variasi ejaan (Budiasa, 2021; Hutauruk et al., 2024; Nisaulfitri & Alamiyah, 2023). Karakteristik tersebut menjadikan data menfess menantang untuk dianalisis secara komputasional, sekaligus bernilai sebagai representasi komunikasi digital informal berbahasa Indonesia. Telegram dipilih sebagai platform karena menyediakan API terbuka yang memudahkan ekstraksi data untuk keperluan penelitian (Peersman et al., 2016).

Text Clustering dan K-Means

Text clustering adalah teknik pengelompokan dokumen berdasarkan kemiripan konten tanpa label kelas, dan merupakan salah satu tugas utama dalam text mining (Aggarwal & Zhai, 2012). K-Means adalah algoritma clustering berbasis partisi yang membagi data ke dalam k cluster dengan meminimalkan jarak intracluster terhadap centroid (MacQueen, 1967). Algoritma ini telah banyak diterapkan pada data teks media sosial, di antaranya untuk pengelompokan ulasan produk e-commerce (Widjaja, 2023) dan analisis konten pemasaran digital (Larasati et al., 2021; Ompo & Pakereng, 2024). Kelebihan utama K-Means adalah efisiensi komputasi dan kemampuannya mengalokasikan seluruh data ke dalam cluster. Namun demikian, K-Means mensyaratkan jumlah cluster k ditentukan di awal, sensitif terhadap inisialisasi centroid, dan mengasumsikan cluster berbentuk konveks dengan ukuran yang relatif seimbang.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah algoritma clustering berbasis kepadatan yang diperkenalkan oleh Ester et al. (1996). DBSCAN mendefinisikan cluster sebagai wilayah padat yang dipisahkan oleh wilayah berkepadatan rendah, dengan dua parameter utama yaitu eps (radius ketetanggaan) dan min_samples (jumlah minimum titik untuk membentuk inti cluster). Keunggulan DBSCAN adalah kemampuannya mendeteksi cluster dengan bentuk arbitrer serta mengidentifikasi data noise secara eksplisit tanpa perlu menentukan jumlah cluster di awal. DBSCAN telah diterapkan untuk deteksi topik dan clustering teks pada data media sosial berbasis kejadian darurat (Huang et al., 2022). Namun, performa DBSCAN sangat bergantung pada pemilihan parameter dan cenderung kurang optimal pada ruang dimensi tinggi akibat fenomena curse of dimensionality, di mana jarak antar titik data menjadi relatif seragam sehingga sulit terbentuk wilayah padat yang bermakna.

Sentence-BERT sebagai Representasi Fitur

Sentence-BERT (SBERT) diperkenalkan oleh Reimers dan Gurevych (2019) sebagai pengembangan

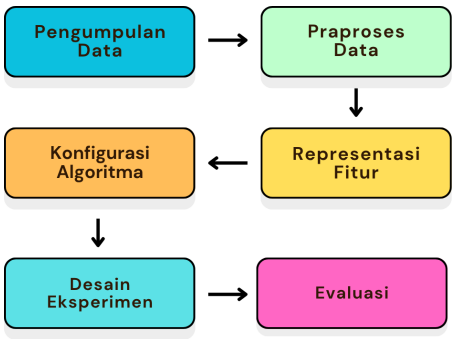
dari model BERT yang dioptimalkan untuk menghasilkan representasi vektor kalimat secara efisien menggunakan arsitektur Siamese network. Berbeda dengan pendekatan berbasis frekuensi kata seperti TF-IDF, SBERT mampu menangkap kemiripan semantik antarkalimat secara lebih efektif, termasuk pada kalimat yang memiliki makna serupa namun ditulis dengan kata yang berbeda. Model paraphrase-multilingual-MiniLM-L12-v2 yang digunakan dalam penelitian ini mendukung lebih dari 50 bahasa termasuk bahasa Indonesia dan menghasilkan vektor berdimensi 384, menjadikannya cocok untuk data menfess yang kaya variasi leksikal. Karena SBERT dirancang untuk memproses teks dalam bentuk yang mendekati aslinya, praproses sebelum embedding dibatasi pada normalisasi unicode dan whitespace serta penghapusan elemen non-linguistik seperti URL dan mention, tanpa normalisasi slang atau penghapusan stopword. Penggunaan cosine similarity sebagai metrik jarak pada ruang embedding SBERT terbukti lebih sesuai dibandingkan Euclidean distance untuk data teks berdimensi tinggi (Reimers & Gurevych, 2019). Untuk teks berbahasa Indonesia, model berbasis BERT yang dilatih khusus seperti IndoBERT (Koto et al., 2020; Wilie et al., 2020) berpotensi memberikan representasi semantik yang lebih akurat karena dilatih pada korpus bahasa Indonesia, dan dapat menjadi alternatif yang dipertimbangkan dalam penelitian lanjutan.

Evaluasi Clustering: Silhouette Score dan Davies-Bouldin Index

Evaluasi clustering tanpa label kelas (internal evaluation) umumnya menggunakan metrik yang mengukur kekompakan dan keterpisahan cluster. Silhouette Score diperkenalkan oleh Rousseeuw (1987) dan mengukur seberapa mirip suatu titik data dengan cluster-nya sendiri dibandingkan dengan cluster terdekat lainnya, dengan rentang nilai –1 hingga 1. Nilai mendekati 1 menunjukkan bahwa titik data berada di cluster yang tepat, nilai mendekati 0 mengindikasikan titik berada di batas antarcluster, dan nilai negatif menandakan kemungkinan salah klaster. Davies-Bouldin Index (DBI) yang diperkenalkan oleh Davies dan Bouldin (1979) mengukur rasio antara penyebaran intracluster dan jarak antarcluster; nilai DBI yang lebih rendah menunjukkan kualitas clustering yang lebih baik. Kedua metrik ini bersifat komplementer dan lazim digunakan bersama dalam penelitian text clustering untuk memberikan gambaran yang lebih komprehensif (Aggarwal & Zhai, 2012).

METODE PENELITIAN

Penelitian ini mempunyai enam tahapan utama yang dilakukan secara berurutan seperti pada Gambar 1 yaitu dimulai dari pengumpulan data, tahap praproses data, representasi fitur, konfigurasi algoritma, dan desain eksperimen, serta evaluasi. Penjelasan lebih detail terkait tahapan dalam penelitian ini dijelaskan di sub bab ini.



Gambar 1. Alur Penelitian

Pengumpulan Data

Data berupa pesan menfess yang diperoleh dari channel Telegram melalui ekstraksi menggunakan skrip Python berbasis Telegram API setelah memperoleh izin pemilik channel. Data disimpan dalam format CSV dengan jumlah awal 42.302 pesan. Contoh data disajikan pada Tabel 1.

Tabel 1. Contoh Dataset Pesan Menfess

Contoh Teks Pesan Menfess
kenapa ya mantan kalau udah dilupain pasti balik lagi, heran deh
OPO WS ORA ISO DI DANDANI, SAYANG MAAF KU TAK BISA LAGI
mls bgt njir sabtu2 brngkt skolahhh😭
aku mau nyari musik di sini tpi yg bisa di simpen gitu, aku lupa namanya ada yg tau
Pengen jadi business woman kaya Khodijah istri nabi

Tabel 1 menampilkan contoh pesan menfess yang diperoleh dari channel Telegram. Pesan-pesan tersebut mencerminkan karakteristik data yang tidak terstruktur, singkat, dan didominasi bahasa informal dengan variasi ejaan, singkatan, serta penggunaan huruf kapital yang tidak konsisten. Keragaman bentuk bahasa ini menjadi tantangan utama dalam proses clustering teks.

Praproses Data dan Skenario

Sebelum clustering, data melewati tahap praproses yang meliputi penghapusan URL, mention, hashtag, dan karakter non-linguistik, normalisasi unicode dan whitespace, serta eliminasi pesan dengan panjang kurang dari 10 karakter, kurang dari 5 kata, dan pesan duplikat. Normalisasi slang dan penghapusan stopword tidak diterapkan karena Sentence-BERT bekerja optimal pada teks yang mendekati bentuk aslinya. Dataset kemudian dibagi ke dalam tiga skenario berdasarkan panjang teks menggunakan K-Means pada fitur word_count: (1) pendek: 0–20 kata, 32.564 data; (2) menengah: 55–128 kata, 1.675 data; dan (3) panjang: 129–283 kata, 314 data.

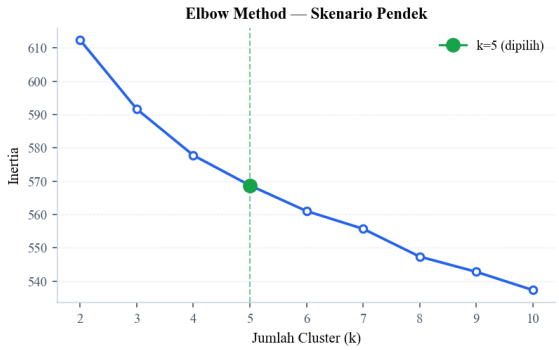
Representasi Fitur

Setiap teks direpresentasikan menggunakan Sentence-BERT embedding dengan model paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2019) yang menghasilkan vektor berdimensi 384. Cosine distance dipilih sebagai metrik jarak karena lebih sesuai untuk vektor embedding dibandingkan Euclidean distance, terutama untuk menghindari dampak curse of dimensionality pada DBSCAN.

Konfigurasi Algoritma

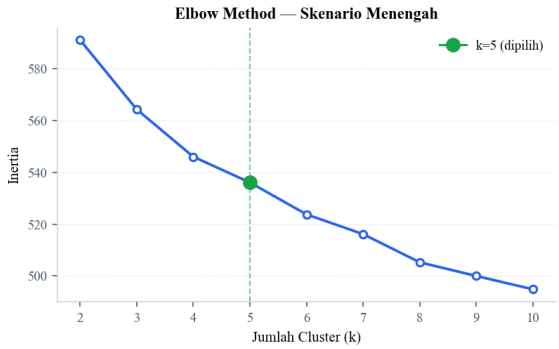
K-Means dikonfigurasi dengan k=5 berdasarkan hasil metode elbow yang mengamati perubahan nilai inertia pada k=2 hingga k=10. Metrik jarak yang digunakan adalah Euclidean distance, karena K-Means secara inheren bekerja dengan menghitung jarak titik data ke centroid untuk memperbarui posisi centroid pada setiap iterasi — operasi yang didefinisikan dalam ruang Euclidean. Meskipun embedding berdimensi tinggi dapat mengurangi efektivitasnya, Euclidean distance pada K-Means tetap lazim digunakan karena algoritma ini tidak bergantung pada definisi wilayah padat, melainkan pada minimisasi jarak intracluster secara global.

DBSCAN dikonfigurasi dengan eps=0,22 dan min_samples=5 berdasarkan analisis k-distance graph. Metrik jarak yang digunakan adalah cosine distance, yang mengukur sudut antarvektor embedding alih-alih besaran absolut jaraknya. Dengan demikian, dua pesan yang memiliki makna serupa akan memiliki cosine distance yang kecil meskipun panjang vektornya berbeda. Cosine distance juga lebih stabil di ruang berdimensi tinggi dibandingkan Euclidean distance yang cenderung menghasilkan jarak antarvektor yang seragam akibat curse of dimensionality (Sun et al., 2020).



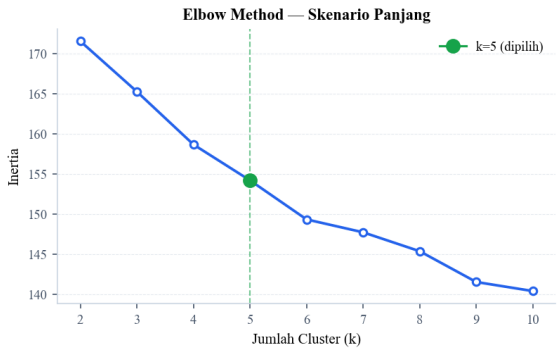
Gambar 2. Elbow Method — Skenario Pendek

Gambar 2 menunjukkan kurva elbow pada skenario teks pendek. Penurunan inertia berlangsung bertahap tanpa titik siku yang tajam, yang merupakan kondisi tipikal pada data teks embedding berdimensi tinggi. Nilai k=5 dipilih pada titik di mana laju penurunan inertia mulai melandai secara relatif, yaitu ketika penambahan cluster tidak lagi memberikan pengurangan inertia yang signifikan.



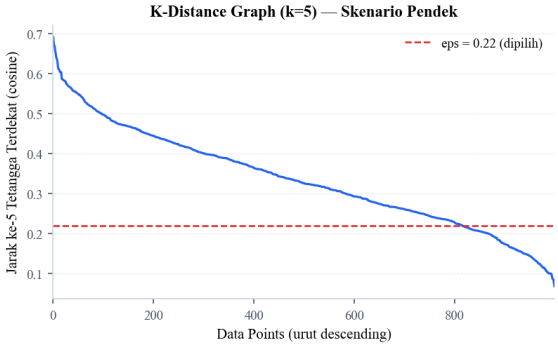
Gambar 3. Elbow Method — Skenario Menengah

Gambar 3 memperlihatkan pola kurva elbow pada skenario teks menengah. Pola penurunan serupa dengan skenario pendek, namun nilai inertia keseluruhan sedikit lebih rendah, mencerminkan distribusi semantik yang relatif lebih terstruktur pada teks dengan panjang menengah. Nilai k=5 tetap dipertahankan untuk menjaga konsistensi perbandingan antarskenario.



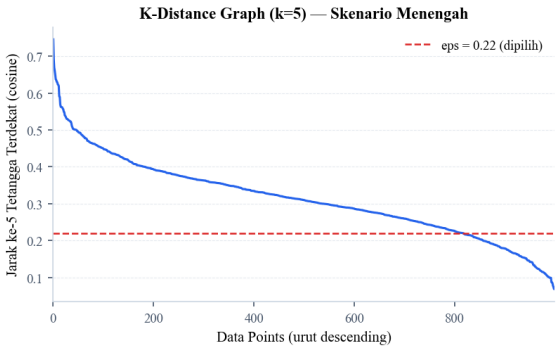
Gambar 4. Elbow Method — Skenario Panjang

Gambar 4 menampilkan kurva elbow pada skenario teks panjang dengan nilai inertia yang secara keseluruhan lebih rendah dibandingkan dua skenario sebelumnya, konsisten dengan jumlah data yang lebih sedikit (300 pesan). Kurva menurun secara gradual dan nilai k=5 dipilih pada titik penurunan yang mulai melandai.



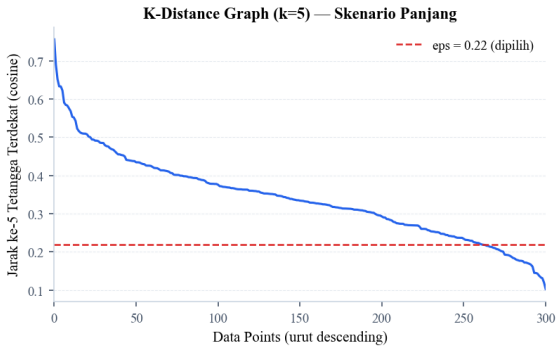
Gambar 5. K-Distance Graph — Skenario Pendek

Gambar 5 menunjukkan k-distance graph pada skenario teks pendek dengan k=5 (sesuai min_samples DBSCAN). Kurva menurun secara gradual tanpa titik siku yang tegas, yang mencerminkan distribusi jarak cosine yang relatif seragam akibat curse of dimensionality pada ruang embedding 384 dimensi. Nilai eps=0,22 dipilih pada area transisi sebelum kurva menurun tajam di bagian akhir, menghasilkan noise rate 71,60%.



Gambar 6. K-Distance Graph — Skenario Menengah

Gambar 6 memperlihatkan k-distance graph skenario menengah dengan pola yang serupa. Tidak adanya titik siku yang jelas mengindikasikan bahwa data teks menengah memiliki distribusi kepadatan yang sangat heterogen di ruang embedding, sehingga DBSCAN dengan eps=0,22 hanya berhasil membentuk 1 cluster dengan noise rate 74,60%.



Gambar 7. K-Distance Graph — Skenario Panjang

Gambar 7 menampilkan k-distance graph skenario panjang. Meskipun jumlah data lebih sedikit (300 pesan), kurva menunjukkan pola yang serupa dengan skenario lainnya. Dengan eps=0,22, DBSCAN menghasilkan 2 cluster dengan noise rate tertinggi (79,67%), namun nilai metrik evaluasi terbaik di antara seluruh skenario, yang mengindikasikan bahwa cluster yang terbentuk bersifat sangat kompak meskipun hanya mencakup sebagian kecil data.

Desain Eksperimen

Dari setiap skenario diambil 1.000 data secara simple random sampling (300 untuk skenario panjang karena keterbatasan data). Desain eksperimen disajikan pada Tabel 2.

Tabel 2. Desain Eksperimen

Aspek	K-Means	DBSCAN
Data (per skenario)	1.000 pesan	1.000 pesan
Representasi fitur	Sentence-BERT (384 dim)	Sentence-BERT (384 dim)
Metrik jarak	Euclidean	Cosine
Parameter utama	k=5	eps=0,22; min_samples=5
Skenario	Pendek, Menengah, Panjang	Pendek, Menengah, Panjang

Tabel 2 merangkum konfigurasi eksperimen yang diterapkan pada kedua metode. K-Means dan DBSCAN menggunakan data dan representasi fitur yang identik, yaitu Sentence-BERT berdimensi 384, sehingga perbedaan hasil yang diperoleh mencerminkan perbedaan karakteristik algoritma, bukan perbedaan input.

Evaluasi

Evaluasi dilakukan secara kuantitatif menggunakan Silhouette Score (Rousseeuw, 1987) dan Davies-Bouldin Index (Davies & Bouldin, 1979). Silhouette Score mengukur kekompakan dan keterpisahan cluster dengan rentang −1 hingga 1, di mana nilai lebih tinggi menunjukkan kualitas cluster yang lebih baik. Davies-Bouldin Index mengukur rasio penyebaran intracluster terhadap jarak antarcluster, di mana nilai lebih rendah menunjukkan hasil yang lebih baik. Khusus pada DBSCAN, kedua metrik dihitung hanya pada data non-noise karena data noise tidak termasuk dalam cluster yang terbentuk.

HASIL DAN PEMBAHASAN

Hasil Evaluasi

Perbandingan hasil evaluasi kuantitatif disajikan pada Tabel 3. (DBI = Davies-Bouldin Index; * dihitung hanya pada data non-noise)

Tabel 3a. Hasil Evaluasi K-Means

Skenario	*k*	Silhouette Score	DBI	Noise Rate
Pendek	5	0,0804	3,8451	0,00%
Menengah	5	0,1403	3,6490	0,00%
Panjang	5	0,0593	3,4552	0,00%

Keterangan: k = jumlah klaster

Tabel 3a menunjukkan hasil evaluasi K-Means pada ketiga skenario. K-Means konsisten membentuk 5 cluster tanpa noise di semua skenario dengan noise rate 0,00%. Silhouette Score tertinggi diperoleh pada skenario menengah (0,1403), mengindikasikan bahwa teks dengan panjang 55–128 kata memiliki distribusi semantik yang relatif lebih terstruktur dibandingkan teks pendek maupun panjang. Nilai Davies-Bouldin Index yang tinggi di seluruh skenario (3,45–3,84) mencerminkan tumpang tindih antarcluster yang signifikan, konsisten dengan heterogenitas topik pada data menfess.

Tabel 3b. Hasil Evaluasi DBSCAN

Skenario	*k*	Silhouette Score	DBI	Noise Rate
Pendek	2	*0,3186**	*1,3714**	71,60%
Menengah	1	*0,0450**	*3,3405**	74,60%
Panjang	2	*0,5492**	*1,1069**	79,67%

Keterangan: k = jumlah klister
** Dihitung hanya pada data non-noise*

Tabel 3b menyajikan hasil evaluasi DBSCAN pada ketiga skenario. DBSCAN menghasilkan Silhouette Score yang lebih tinggi dan Davies-Bouldin Index yang lebih rendah dibandingkan K-Means pada skenario pendek dan panjang, namun hasil tersebut dihitung hanya pada data non-noise. Perlu dicatat bahwa noise rate yang sangat tinggi di seluruh skenario (71,60%–79,67%) menunjukkan bahwa sebagian besar data tidak berhasil dikelompokkan, sehingga perbandingan metrik dengan K-Means perlu diinterpretasikan secara hati-hati. Pada skenario menengah, DBSCAN hanya membentuk 1 cluster, mengindikasikan kegagalan dalam memisahkan topik secara bermakna.

Tabel 4. Ringkasan Cluster K-Means — Skenario Pendek

C	n	Contoh pesan
0	113	ih apalah tadi ketemu mantan tercinta jadi gamon lagi kan
1	284	ayo dong temenan sama aku
2	212	anjeng lah ternyata selama ini cuma jadi pelampiasan badjingan emng
3	239	tidak ketemu kangen ketemu makin kangen serumah saja tidak si kita
4	152	plis ajarin gue freelance yang gratis gue mau ngasilin duit biar tidak beban bngt jdi anak

Keterangan: C = label/urutan klaster; n = total data

Tabel 4 menyajikan contoh pesan representatif dari masing-masing cluster K-Means pada skenario teks pendek. Kelima cluster menunjukkan variasi topik yang dapat diidentifikasi dari contoh pesan, meliputi percintaan dan hubungan romantis (cluster 0), interaksi dan chat menfess (cluster 1), ekspresi perasaan ringan (cluster 2), interaksi sosial dan pertemanan (cluster 3), serta kebutuhan praktis dan informasi (cluster 4).

Tabel 5. Ringkasan Cluster K-Means — Skenario Menengah

C	n	Contoh pesan
0	238	crush gue kenapa malu plis kalo ketemu gue trs kaya grogi gtu di ajak ngobrol juga diem gamau jawab gue kata temennya padahal suka
1	307	gw kan dri h 1 ramadhan ni tidak puasa ya tapi tu 1 rumah puasa cuy gw mau minum mau makan ya gaenak dong jdi kya ikutan puasa

2	215	kenapa ya ortu cm liat sisi yang kurang saja misal gw makan banyak bet yang diliat pas tggl dikit trus ngomong makan yang banyak dikit bnget nasinya
3	101	critnya gua di putusin awal nya sosmed gua tidak di blok sama dia ada sedikit gelisah gtau knp tpi gua msi bisa fine tpi skrng gua uda tidak bisa tau kbr dia
4	139	capek nangis orang yang udah pada keterima bahkan dapet kampus tapi gue masih berjuang buat mandiri dan tiap ortu dapet pengumuman kalau temen gue keterima

Keterangan: C = label/urutan klaster; n = total data

Tabel 5 menampilkan contoh pesan dari cluster K-Means pada skenario teks menengah. Dengan panjang teks yang lebih substansial, pesan-pesan dalam tiap cluster cenderung lebih naratif dan memuat konteks yang lebih kaya, sehingga perbedaan antarcluster relatif lebih terlihat dibandingkan skenario pendek. Topik yang dapat diidentifikasi mencakup hubungan romantis (cluster 0), cerita harian dan pertemanan (cluster 1), kondisi keluarga dan ekonomi (cluster 2), konflik dan emosi (cluster 3), serta kehidupan sekolah dan perkuliahan (cluster 4).

Tabel 6. Ringkasan Cluster K-Means — Skenario Panjang

C	n	Contoh pesan
0	91	aku mau sedikit crita nih kan aku lagi deket sama cowo temen sekelas nih aku deket sama dia itu gara di comblangin sama temen sekelas nah aku denger dari orang
1	68	guys aku kan pkl dptnya sama cowo ya berdua doang jir btw ni klmpk udh ditentuin sama guru jadi tidak bisa milih trs si cwo ini punya cewe kmrn aku ngechat
2	54	hai gue mau curhat dikit jadi gue adalah anak yang terlahir di keluarga yang serba kekurangan bahkan motor saja tidak punya ampe keluarga aku pinjem punya nenek
3	55	guys maaf emang salah ya kalau aku sering cerita tentang mantan aku ngungkit yang uda berlalu aku tidak gamon aku cuma mau ngeluarin isi hati aku
4	32	mnfes ah taik tidak bisa kirim foto yauda gini cerita nya awl ny tuh aku tuh iseng main leo karna bot anon aku tuh tidak ada jadi aku main leo

Keterangan: C = label/urutan klaster; n = total data

Tabel 6 menyajikan contoh pesan dari cluster K-Means pada skenario teks panjang. Pesan-pesan pada skenario ini bersifat naratif dan panjang, sehingga cluster yang terbentuk cenderung mencerminkan gaya penceritaan dibandingkan topik tunggal. Cluster yang dapat diidentifikasi meliputi cerita percintaan panjang (cluster 0), log percakapan berbasis chat (cluster 1), cerita keluarga dan kondisi ekonomi (cluster 2), refleksi dan ekspresi emosional (cluster 3), serta interaksi online dan penggunaan bot (cluster 4).

Tabel 7. Ringkasan Cluster DBSCAN — Skenario Pendek (noise: 716 pesan)

C	n	Contoh pesan
0	276	gw kelai sama pcr gw tpi gw kangen dia
1	8	td udah ngantuk tapi skrng mlh ilang ngntuknya gegara pegang hp

Keterangan: C = label/urutan klaster; n = total data

Tabel 7 menampilkan dua cluster yang berhasil dibentuk DBSCAN pada skenario teks pendek dari 284 data non-noise. Cluster 0 yang jauh lebih besar (276 pesan) mengelompokkan pesan bertema perasaan dan percintaan, sedangkan cluster 1 yang sangat kecil (8 pesan) berisi pesan bertema ngantuk dan aktivitas malam hari. Ketimpangan ukuran cluster ini menunjukkan bahwa hanya sebagian kecil pesan yang memiliki kedekatan semantik cukup tinggi untuk membentuk cluster tambahan di luar kelompok utama.

Tabel 8. Ringkasan Cluster DBSCAN — Skenario Menengah (noise: 746 pesan)

C	n	Contoh pesan
0	254	dia cpe krna aku pengen di ngertiin trs sma dia sdngkan aku tidak ngertiin dia samsek tpi sblmnya aku udh prnh bilng ke dia mw nya gmna dia slalu jawab gatau

Keterangan: C = label/urutan klaster; n = total data

Tabel 8 menunjukkan bahwa DBSCAN hanya berhasil membentuk satu cluster pada skenario teks menengah, yang berisi 254 pesan dari 1.000 data (noise rate 74,60%). Cluster tunggal ini mengelompokkan pesan bertema cerita dan curhatan umum, sementara 746 pesan lainnya diklasifikasikan sebagai noise. Kondisi ini mengindikasikan bahwa distribusi semantik teks menengah terlalu heterogen untuk memungkinkan DBSCAN membentuk lebih dari satu wilayah padat yang bermakna.

Tabel 9. Ringkasan Cluster DBSCAN — Skenario Panjang (noise: 239 pesan)

C	n	Contoh pesan
0	48	guys aku kan pkl dptnya sama cowo ya berdua doang jir btw ni klmpk udh ditentuin sama guru jadi tidak bisa milih trs si cwo ini punya cewe kmrn aku ngechat
1	13	mau sharing saja gua kan gua punya kenalan nih di tele karna main bot gtu trus tidak lama setelah itu kita ada komitmen gtu trus harusnya sabtu kemaren kita ...

Keterangan: C = label/urutan klaster; n = total data

Tabel 9 menyajikan dua cluster yang dibentuk DBSCAN pada skenario teks panjang dari 61 data non-noise. Cluster 0 (48 pesan) mengelompokkan narasi interaksi dan percakapan berbasis chat WA, sedangkan cluster 1 (13 pesan) berisi pengalaman emosional personal yang lebih intim. Meskipun jumlah data yang masuk cluster sangat sedikit, kedekatan semantik yang tinggi dalam cluster ini menghasilkan nilai Silhouette Score terbaik (0,5492) di antara seluruh kondisi eksperimen.

Secara umum, DBSCAN menghasilkan Silhouette Score lebih tinggi dan DBI lebih rendah dibandingkan K-Means pada skenario pendek dan panjang. Namun, metrik tersebut hanya dihitung pada data non-noise dengan proporsi noise sangat tinggi (71,60%–79,67%). Pada skenario menengah, DBSCAN hanya membentuk 1 cluster sehingga kemampuan pemisahan topik tidak bermakna.

Analisis K-Means

K-Means mengalokasikan seluruh data ke dalam 5 cluster tanpa noise di semua skenario. Nilai Silhouette Score yang rendah (0,0593–0,1403) dan DBI yang tinggi (3,4552–3,8451) mengindikasikan batas antarcluster yang masih tumpang tindih, sejalan dengan karakteristik menfess yang topiknya sangat beragam (Widjaja, 2023).

Analisis DBSCAN

DBSCAN menghasilkan karakteristik yang sangat berbeda antarketiga skenario (Tabel 7–9). Pada skenario pendek terbentuk 2 cluster: cluster besar bertema perasaan dan percintaan (276 pesan) dan cluster kecil bertema ngantuk dan malam hari (8 pesan), dengan noise rate 71,60%. Pada skenario menengah hanya terbentuk 1 cluster bertema cerita dan curhatan umum (254 pesan) dengan noise rate 74,60%, menunjukkan DBSCAN tidak mampu memisahkan topik secara bermakna. Pada skenario panjang terbentuk 2 cluster yang lebih spesifik: narasi interaksi dan chat WA (48 pesan) serta pengalaman emosional personal (13 pesan), dengan noise rate tertinggi 79,67%. Tingginya noise di seluruh skenario berkaitan dengan heterogenitas data menfess dan curse of dimensionality pada ruang 384 dimensi, fenomena serupa yang dilaporkan oleh Huang et al. (2022).

Pengaruh Panjang Teks

Pembagian dataset berdasarkan panjang teks memiliki dua implikasi mendasar terhadap hasil penelitian ini. Pertama, pembagian skenario merupakan syarat struktural agar DBSCAN dapat membentuk cluster sama sekali. Tanpa pembagian ini, data yang mencakup rentang 0–283 kata akan jauh lebih heterogen secara semantik, sehingga distribusi vektor embedding menjadi sangat tersebar dan DBSCAN hampir tidak dapat menemukan wilayah padat yang bermakna — noise rate diperkirakan akan mendekati atau melebihi 90%. Dengan kata lain, pembagian skenario bukan sekadar variasi eksperimen, melainkan pra-pemrosesan implisit yang secara struktural menguntungkan DBSCAN.

Kedua, meskipun pembagian skenario memungkinkan DBSCAN bekerja, noise rate tetap sangat tinggi di seluruh skenario (71,60%–79,67%). Akibatnya, metrik evaluasi hanya mencerminkan kualitas cluster pada sebagian kecil data yang paling homogen, bukan representasi keseluruhan dataset. Ini berarti keunggulan angka metrik DBSCAN merupakan hasil dari dua faktor yang saling memperkuat: homogenitas data yang diciptakan oleh pembagian skenario, dan selektivitas inheren DBSCAN yang hanya mengevaluasi data non-noise. K-Means, sebaliknya, tidak diuntungkan oleh mekanisme serupa — seluruh data dievaluasi tanpa pengecualian, sehingga perbandingan metrik antarkedua metode perlu diinterpretasikan dengan mempertimbangkan perbedaan mendasar ini.

Implikasi Pemilihan Metode

Jika tujuan analisis menekankan kualitas dan kekompakan cluster, DBSCAN dapat dipilih dengan konsekuensi tingginya data tidak terklasifikasi. Sebaliknya, jika tujuan adalah pengelompokan menyeluruh untuk gambaran umum topik di berbagai skenario panjang teks, K-Means lebih sesuai karena performa yang lebih stabil dan konsisten.

KESIMPULAN

Penelitian ini membandingkan K-Means dan DBSCAN untuk clustering teks pesan menfess di Telegram menggunakan Sentence-BERT embedding pada tiga skenario panjang teks. Hasil menunjukkan bahwa pada skenario pendek, DBSCAN menghasilkan Silhouette Score 0,3186 dan DBI 1,3714 (noise rate 71,60%), lebih baik dibandingkan K-Means dengan 0,0804 dan 3,8451. Pada skenario menengah, K-Means (0,1403; 3,6490) mengungguli DBSCAN yang hanya membentuk 1 cluster dengan noise rate 74,60%. Pada skenario panjang, DBSCAN mencapai hasil terbaik (Silhouette 0,5492; DBI 1,1069) namun dengan noise rate 79,67%.

K-Means lebih efektif untuk pengelompokan data secara menyeluruh di seluruh skenario, sedangkan DBSCAN unggul secara metrik pada teks pendek dan panjang namun menghasilkan proporsi noise yang sangat tinggi. Penelitian selanjutnya disarankan untuk mengeksplorasi reduksi dimensi (UMAP), algoritma yang lebih adaptif (HDBSCAN), kombinasi dengan topic modeling (BERTopic), serta model embedding berbasis bahasa Indonesia (IndoBERT).

DAFTAR PUSTAKA

Adawiyah, R. (2023). Cluster text random opinion tweet in Yogyakarta using automatic clustering. *Jurnal Penelitian Rumpun Ilmu Teknik*, *2*(1), 73–89.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 77–128). Springer. https://doi.org/10.1007/978-1-4614-3223-4_4

Budiasa, I. G. (2021). Slang language in Indonesian social media. *Lingual: Journal of Language and Culture*, *11*(1). https://doi.org/10.24843/ljlc.2021.v11.i01.p06

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

Hidayat, R., Id, I. D., Amanda, F. P., Mumtaza, D. A., Maulana, M. A., Tamir, A. A., & Firmansyah, I. (2024). Kajian opini publik terhadap tayangan Clash of Champions menggunakan Support Vector Machine. *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, *4*(2), 274–281. https://doi.org/10.46880/tamika.Vol4No2.pp274-281

Huang, L., Shi, P., & Zhu, H. (2022). Early detection of emergency events from social media: A new text clustering approach. *Natural Hazards*, *111*(3), 2387–2409. https://doi.org/10.1007/s11069-021-05081-1

Hutauruk, B., Agatha, Manurung, O., Sinaga, S., & Aryani, N. (2024). Slang language in social media X (Twitter). *Journal of English Language Teaching and Linguistics in Applied Linguistics*, *2*(1). https://doi.org/10.69820/jeltlal.v2i1.125

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Kurniawan, M., & Achmadi, H. (2024). Sentiment analysis and clustering of ISP service users based on social media platform X in Indonesia using K-Means method. *TRANSEKONOMIKA: Akuntansi, Bisnis Dan Keuangan*, *4*(6), 1096–1104. https://doi.org/10.55047/transekonomika.v4i6.728

Larasati, A., Maren, R., & Wulandari, R. (2021). Utilizing elbow method for text clustering optimization in analyzing social media marketing content of Indonesian e-commerce. *Jurnal Teknik Industri*, *23*(2), 111–120. https://doi.org/10.9744/jti.23.2.111-120

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

Nisaulfitri, N. D., & Alamiyah, S. S. (2023). Komunikasi hyperpersonal dalam chatting anonim pengguna bot anonymous chat di Telegram. *JIIP: Jurnal Ilmiah Ilmu Pendidikan*, *6*(11), 8654–8662. https://doi.org/10.54371/jiip.v6i11.3182

Ompo, F. A. A., & Pakereng, M. A. I. (2024). Penerapan text mining untuk advertising pada data tweets Zalora Indonesia dengan menggunakan metode K-Means clustering. *Progresif: Jurnal Ilmiah Komputer*, *20*(1). https://doi.org/10.35889/progresif.v20i1.1576

Peersman, C., Daelemans, W., & van Vaerenbergh, L. (2016). Predicting age and gender in online social networks. *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, 37–44. https://doi.org/10.1145/2065023.2065035

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://aclanthology.org/D19-1410/

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sun, X., Qin, Y., Dong, L., & Xu, T. (2020). High-dimensional text clustering by dimensionality reduction and improved density peak. *Wireless Communications and Mobile Computing*, *2020*, 8881112. https://doi.org/10.1155/2020/8881112

Widjaja, A. E. (2023). Text mining application with K-Means clustering to identify sentiments and popular topics: A case study of the three largest online marketplaces in Indonesia. *Journal of Applied Data Sciences*, *4*(4). https://doi.org/10.47738/jads.v4i4.134

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2020)*, 843–857. https://aclanthology.org/2020.aacl-main.85

Zikirlah, H. A., Paula, I., Fazilla, M., Annisa, R., & Fitriana, L. A. (2025). Perbandingan kinerja Naïve Bayes, Support Vector Machine, dan K-Nearest Neighbor dalam analisis sentimen Mobile Legends. *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, *5*(2), 228–235.