Tersedia Online : http://e-journals.unmul.ac.id/

ADOPSI TEKNOLOGI DAN SISTEM INFORMASI (ATASI)

Alamat Jurnal : http://e-journals2.unmul.ac.id/index.php/atasi/index

Implementasi *Cosine Similarity* dan *K-Means Clustering* pada Data Kuesioner untuk Analisis Efektivitas Pembelajaran Daring

Herlina [1)*]
[1)]Program Studi Informatika, Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta

E-Mail : herlina@uajy.ac.id [1)]

ARTICLE INFO

Article history:
Received : 01-08-2025
Revised : 13-08-2025
Accepted : 06-11-2025
Available online : 29-04-2026

Keywords:
online learning
lecture obstacles
cosine similarity
K-Means clustering
unsupervised learning

ABSTRACT

This study aims to evaluate student's readiness in online learning by utilizing the K-Means clustering method to group survey results based on data similarity. K-Means, as an unsupervised learning method, functions to classify data into several clusters, so that hidden patterns relevant to online learning readiness can be identified. In this study, text data obtained from student surveys were processed through a pre-processing stage to improve the accuracy and efficiency of the algorithm in grouping data. In addition, cosine similarity was applied to measure the level of similarity between texts, making it easier to identify similarities in obstacles faced by students during the online learning process. Clustering analysis shows that some of the main obstacles faced by students in online learning include inadequate internet signal, less supportive learning environment, difficulty in understanding the material, and communication barriers both with lecturers and in group collaboration. These results provide insight for universities to develop more adaptive online learning strategies, by addressing specific obstacles faced by students. Thus, the quality and effectiveness of the teaching and learning process can be significantly improved, helping students optimize their online learning experience.

Kata Kunci :
pembelajaran daring
kendala perkuliahan
cosine similarity
K-Means clustering
unsupervised learning

ABSTRAK

Penelitian ini bertujuan untuk mengevaluasi kesiapan mahasiswa dalam pembelajaran daring dengan memanfaatkan metode *K-Means clustering* untuk mengelompokkan hasil survei berdasarkan kemiripan data. *K-Means*, sebagai metode *unsupervised learning*, berfungsi mengklasifikasikan data ke dalam beberapa cluster, sehingga pola tersembunyi yang relevan dengan kesiapan pembelajaran daring dapat diidentifikasi. Dalam penelitian ini, data teks yang diperoleh dari survei mahasiswa diproses melalui tahap *pre-processing* untuk meningkatkan akurasi dan efisiensi algoritma dalam mengelompokkan data. Selain itu, *cosine similarity* diterapkan untuk mengukur tingkat kemiripan antar teks, sehingga mempermudah dalam mengidentifikasi kesamaan kendala yang dihadapi mahasiswa selama proses pembelajaran daring. Analisis clustering menunjukkan bahwa beberapa kendala utama yang dihadapi mahasiswa dalam pembelajaran daring meliputi sinyal internet yang tidak memadai, lingkungan belajar yang kurang mendukung, kesulitan dalam memahami materi, serta hambatan komunikasi baik dengan dosen maupun dalam kolaborasi kelompok. Hasil ini memberikan wawasan bagi pihak universitas untuk mengembangkan strategi pembelajaran daring yang lebih adaptif, dengan upaya mengatasi kendala-kendala spesifik yang dihadapi mahasiswa. Dengan demikian, kualitas dan efektivitas proses belajar-mengajar dapat ditingkatkan secara signifikan, membantu mahasiswa dalam mengoptimalkan pengalaman belajar daring.

*) Corresponding Author
https://doi.org/10.30872/atasi.v5i1.3577

1. PENDAHULUAN

Pandemi Covid-19 telah membawa perubahan yang signifikan pada berbagai aspek kehidupan, termasuk dalam proses pembelajaran. Sebelum pandemi, pembelajaran umumnya dilakukan secara tatap muka, namun kini telah beralih ke metode daring. Transisi ini menimbulkan berbagai tantangan baru, terutama dalam hal efektivitas pembelajaran. Salah satu permasalahan utama yang dihadapi mahasiswa adalah sulitnya mempertahankan fokus selama proses belajar daring akibat berbagai distraksi yang muncul di lingkungan rumah (Dhawan, 2020). Tidak hanya mahasiswa, dosen pun menghadapi kesulitan dalam memantau aktivitas belajar mahasiswa secara individu, karena pada saat yang sama mereka harus berkonsentrasi pada penyampaian materi perkuliahan (Adedoyin & Soykan, 2023). Meskipun demikian, seluruh pihak berharap bahwa perubahan media pembelajaran dari luring ke daring tidak akan mempengaruhi kualitas hasil belajar mahasiswa yang dihasilkan oleh perguruan tinggi. Berbagai studi menunjukkan bahwa kualitas pembelajaran daring bergantung pada faktor, seperti ketersediaan teknologi, kesiapan dosen dan mahasiswa, serta efektivitas metode pembelajaran yang diterapkan.

Setiap mahasiswa memiliki tingkat kesiapan yang berbeda dalam mengikuti perkuliahan daring, karena berbagai kendala yang dihadapi. Faktor-faktor seperti kualitas koneksi internet, ketersediaan perangkat (komputer atau ponsel), kondisi lingkungan, serta dukungan keluarga sering kali menjadi penghambat utama dalam proses pembelajaran daring. Tidak dapat dipungkiri bahwa beberapa mahasiswa mungkin menghadapi hambatan serupa (Abbas et al., 2017; Adnan, 2020). Oleh karena itu, penting untuk mengelompokkan kendala-kendala tersebut ke dalam kategori-kategori tertentu, yang masing-masing diberi label yang mewakili masalah yang ada. Dengan adanya kategori ini, akan lebih mudah untuk menemukan solusi yang efektif dan efisien, sehingga proses pembelajaran daring dapat berjalan lebih optimal. Setiap kategori mencerminkan permasalahan utama yang perlu diatasi. Namun, tantangan terbesar dalam proses ini adalah kompleksitas dan volume data yang sangat besar, terutama jika melibatkan ribuan mahasiswa. Setiap data memiliki berbagai parameter yang harus dianalisis, sehingga memerlukan waktu dan sumber daya yang signifikan.

Penelitian ini menganalisis tingkat kesiapan mahasiswa dalam mengikuti pembelajaran daring berdasarkan data hasil survei. Analisis dilakukan menggunakan metode *K-Means Clustering*, yaitu metode yang mengelompokkan (*clustering*) data yang memiliki kesamaan ke dalam kelompok secara mandiri (*unsupervised*). *K-Means* merupakan salah satu metode klasterisasi berbasis *unsupervised learning* yang berfungsi untuk mengelompokkan data secara otomatis ke dalam sejumlah kelompok (*k-cluster*) sesuai dengan nilai *k* yang telah ditentukan sebelumnya. Metode ini bekerja dengan cara meminimalkan jarak antar data dalam satu kelompok dan memaksimalkan jarak antar kelompok yang berbeda. *K-Means* dikenal karena efisiensinya dalam menangani dataset besar dan kemampuannya mengidentifikasi pola tersembunyi dalam data (Münz dkk., 2007). Metode ini banyak diterapkan dalam berbagai bidang, termasuk analisis kesiapan mahasiswa dalam pembelajaran daring (Kumar Yadav, 2012; Singh et al., 2016), serta pengolahan teks dan citra dalam berbagai aplikasi lainnya (Bustami et al., 2023).

Implementasi metode *K-Means Clustering* pada penelitian ini diharapkan dapat membantu universitas dalam menganalisis data survei untuk mengevaluasi efektivitas pembelajaran daring. Dengan pengelompokan otomatis berdasarkan pola kesamaan, universitas dapat lebih mudah mengidentifikasi kelompok mahasiswa dengan masalah yang serupa. Jumlah data yang besar tidak lagi menjadi hambatan karena pengolahan dilakukan oleh komputer. Hasil analisis ini juga diharapkan memberikan wawasan berharga bagi universitas untuk merancang strategi peningkatan pembelajaran daring, memungkinkan deteksi masalah yang lebih cepat dan penanganan yang lebih tepat waktu. Selain itu, penelitian ini membuka peluang untuk memprediksi tren atau pola kendala di masa mendatang, memungkinkan langkah antisipatif yang lebih dini. Dengan pendekatan proaktif ini, universitas tidak hanya merespon masalah yang ada, tetapi juga memastikan kelancaran pembelajaran daring di masa depan, sehingga proses belajar-mengajar dapat berjalan lebih efisien dan efektif.

2. TINJAUAN PUSAKA

Perkembangan teknologi digital telah mendorong transformasi dalam sistem pembelajaran, terutama saat pandemi Covid-19 yang memaksa perguruan tinggi beralih dari pembelajaran tatap muka ke daring. Perubahan ini tidak hanya memunculkan tantangan teknis tetapi juga sosial dan psikologis, baik bagi dosen maupun mahasiswa (Adedoyin & Soykan, 2023; Dhawan, 2020). Berbagai studi menunjukkan bahwa efektivitas pembelajaran daring sangat bergantung pada kesiapan infrastruktur, kompetensi digital, dan dukungan lingkungan belajar. Mahasiswa kerap menghadapi kesulitan dalam menjaga konsentrasi dan memahami materi, sementara dosen mengalami keterbatasan dalam melakukan pemantauan dan interaksi personal.

Seiring meningkatnya kebutuhan untuk mengevaluasi proses pembelajaran daring, pendekatan berbasis data seperti analisis kuesioner menjadi metode yang banyak digunakan. Namun, kompleksitas dan volume data kuesioner, terutama yang berbentuk teks, menyulitkan analisis manual. Oleh karena itu, teknik klasterisasi seperti *K-Means* digunakan untuk mengelompokkan data berdasarkan kemiripan isi, sehingga dapat diidentifikasi pola-pola utama dalam pengalaman mahasiswa (Münz et al., 2007; Singh et al., 2016). *K-Means* adalah algoritma *unsupervised learning* yang terbukti efisien untuk mengelompokkan data besar dan mengenali keterkaitan antar data berdasarkan fitur tertentu.

Untuk mendukung proses klasterisasi, metode pengukuran kemiripan teks seperti *cosine similarity* digunakan. *Cosine similarity* mampu menghitung tingkat kesamaan antar dokumen berbasis sudut vektor dalam ruang multidimensi (Kumar et al., 2020). Dalam konteks data kuesioner, metode ini memungkinkan

https://doi.org/10.30872/atasi.v5i1.3577

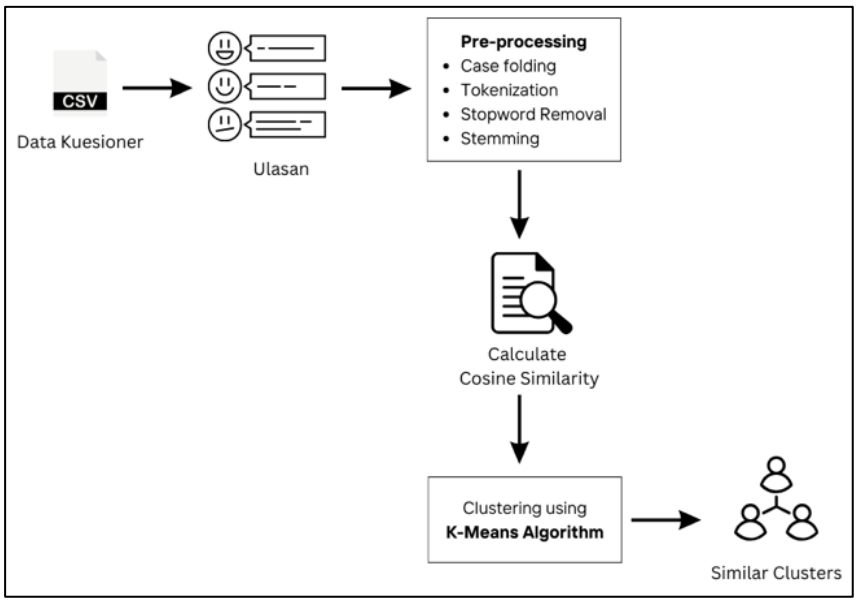
pengelompokan respons mahasiswa yang memiliki kata atau frasa serupa, seperti keluhan mengenai sinyal internet atau keterbatasan komunikasi dengan dosen. Pendekatan ini meningkatkan akurasi dan relevansi hasil klasterisasi karena mempertimbangkan semantik dari konten teks.

Sebelum dilakukan pengukuran kemiripan dan klasterisasi, data kuesioner perlu melalui tahap *pre-processing*. Tahapan ini mencakup proses seperti *case folding*, *tokenization*, *stopword removal*, dan *stemming* untuk menyederhanakan dan membersihkan teks dari elemen yang tidak penting (Gupta & Lehal, 2009; Jianqiang & Xiaolin, 2017). Dalam penelitian yang menggunakan teks berbahasa Indonesia, tools seperti Sastrawi digunakan untuk pengolahan kata dasar. Proses ini penting agar algoritma pemrosesan teks dapat bekerja secara optimal, meningkatkan efisiensi komputasi dan hasil klasifikasi data.

Implementasi kombinasi cosine similarity dan K-Means clustering pada data kuesioner pembelajaran daring terbukti efektif dalam mengelompokkan kendala mahasiswa ke dalam lima kategori utama, seperti masalah sinyal, lingkungan belajar yang tidak mendukung, serta komunikasi dan kolaborasi yang sulit (Arora et al., 2016; Bustami et al., 2023). Dengan hasil klasterisasi tersebut, institusi pendidikan dapat merancang strategi yang lebih adaptif dan tepat sasaran dalam meningkatkan kualitas pembelajaran daring. Metode ini juga dapat dijadikan sebagai pendekatan prediktif untuk memetakan dan mengantisipasi tantangan serupa di masa depan.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini berasal dari kuesioner. Kuesioner berisi pertanyaan terkait aktivitas dan pengalaman mahasiswa selama mengikuti perkuliahan daring, khususnya mengenai permasalahan atau kendala yang sering dialami. Analisis data kuesioner dilakukan melalui beberapa tahap, yaitu: (1) *pre-processing* data, (2) pengukuran kemiripan data kuesioner, (3) pembentukan *cluster*, dan (4) evaluasi hasil *cluster*.



Gambar 1. Tahapan Penelitian

Dalam penelitian ini, tahap *pre-processing* sebagai metode pemrosesan data untuk memastikan bahwa data masukan memiliki kualitas yang sesuai dengan kebutuhan model. Proses ini bertujuan meminimalkan informasi yang tidak relevan, mengurangi *noise*, serta menyederhanakan representasi data sehingga pemrosesan dapat berlangsung secara optimal (Jianqiang & Xiaolin, 2017). Gambar 1 menunjukkan proses yang dilakukan pada tahap pre-processing. Tahapan yang digunakan meliputi *case folding* untuk mengonversi semua huruf menjadi huruf kecil dan menghapus karakter non-alfanumerik, *tokenization* untuk memecah kalimat menjadi unit kata atau *token*, *filtering* untuk memilih kata yang relevan, *stopword removal* untuk menghilangkan kata umum yang tidak memiliki makna spesifik, serta *stemming* untuk mengembalikan kata berimbuhan ke bentuk dasar. Implementasi teknik ini secara langsung memengaruhi kinerja algoritma pembelajaran mesin yang digunakan dalam penelitian, karena mampu meningkatkan efisiensi komputasi, akurasi model, dan kemampuan model dalam mengenali pola signifikan pada data teks. Dengan demikian, *pre-processing* yang tepat berperan strategis dalam menurunkan kompleksitas data, mempercepat waktu pemrosesan, dan meningkatkan kualitas hasil analisis teks (Gupta & Lehal, 2009).

Data yang telah melalui tahap *pre-processing* kemudian diukur tingkat kemiripannya menggunakan *cosine similarity*. Metode ini menghitung kemiripan berbasis arah vektor dan sering digunakan dalam pengolahan bahasa alami untuk menilai kesamaan antar dokumen atau kata. Rumus *cosine similarity* dapat dilihat pada Pers. (1). Nilai *cosine similarity* berkisar antara 0 hingga 1, di mana nilai mendekati 1 menunjukkan kemiripan yang tinggi, sedangkan nilai mendekati 0 menunjukkan perbedaan yang signifikan (Kumar et al., 2020). Hasil perhitungan ini menjadi dasar pembentukan *cluster*.

https://doi.org/10.30872/atasi.v5i1.3577

$$cosine\_similarity\,(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \times \sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad \text{.......................... (1)}$$

Tahap berikutnya adalah melakukan *clustering* menggunakan algoritma *K-Means*. Algoritma ini termasuk metode *unsupervised learning* yang membagi data ke dalam *k cluster* berdasarkan jarak terhadap pusat *cluster* (*centroid*) (Ikotun et al., 2023). Proses dimulai dengan menentukan jumlah *cluster* (*k*) yang akan digunakan, kemudian memilih *centroid* awal secara acak. Pada penelitian ini, jarak antar data dihitung berdasarkan hasil *cosine similarity* pada Pers. (1), yang dikonversi menjadi *cosine distance* dengan menggunakan Pers. (2).

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - c_{jm})^2} \quad \text{.......................... (2)}$$

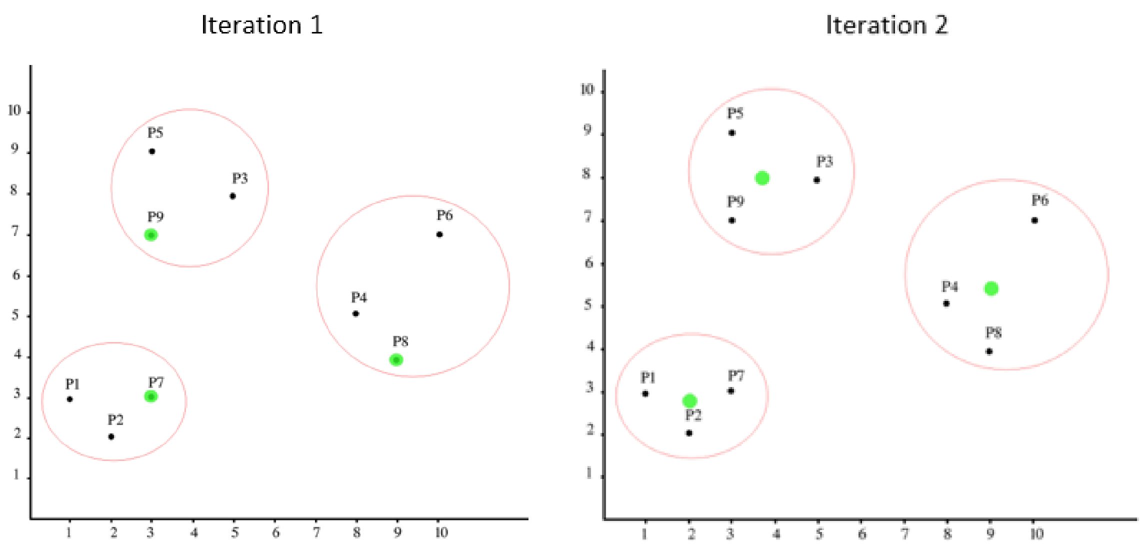
Setiap data ditempatkan pada *cluster* dengan jarak terkecil terhadap *centroid*, kemudian *centroid* diperbarui sebagai rata-rata seluruh anggota *cluster* dengan menggunakan Pers. (3).

$$c_j = \frac{1}{|c_j|} \sum_{x_i \in C_j} x_i \quad \text{.......................... (3)}$$

Proses ini diulang hingga centroid tidak berubah signifikan atau iterasi maksimum tercapai. Jumlah *k* optimal ditentukan menggunakan *Silhouette score* (Kumar Sutrakar & Mogre, 2025). Pers. (4) menunjukkan rumus yang digunakan untuk menghitung *Silhouette score* dengan $a(i)$ adalah jarak rata-rata antara data *i* dan anggota *cluster*-nya, $b(i)$ adalah jarak rata-rata antara data *i* dan cluster terdekat. Nilai *Silhouette* berkisar -1 hingga 1; nilai mendekati 1 menunjukkan pemisahan *cluster* yang baik.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad \text{.......................... (4)}$$

Langkah pada algoritma *K-Means clustering* ini akan dilakukan dalam beberapa iterasi. Seperti yang ditunjukkan pada Gambar 2 (Dharma, 2019), terdapat perubahan nilai *centroid* pada setiap iterasi. Perubahan nilai *centroid* dapat mempengaruhi data yang dikelompokkan di dalam *cluster* tersebut. Langkah ini dilakukan berulang sampai dengan tidak ada perubahan lagi pada nilai *centroid* pada setiap *k*.



Gambar 2. Algoritma *K-Means Clustering*

4. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang digunakan dalam penelitian berasal dari isian kuesioner. Responden yang menjadi target dalam pengisian kuesioner ini adalah mahasiswa yang pernah mengikuti perkuliahan atau proses belajar mengajar secara daring. Sebanyak 323 responden berpartisipasi dalam pengisian kuesioner, terdiri dari 251 laki-laki dan 72 perempuan. Dari total 323 responden, 39,94% berasal dari luar Pulau Jawa, sementara 60,06% berasal dari Pulau Jawa.

https://doi.org/10.30872/atasi.v5i1.3577

Tabel 1. Contoh Data Kuesioner: Profil Responden

ID	Tahun	Jenis Kelamin	IPK	Domisili	Perangkat
1	2018	Laki-laki	3,09	Yogyakarta	Laptop/Handphone
2	2018	Laki-laki	3,40	Yogyakarta	Laptop/Handphone
3	2019	Laki-laki	3,81	Makassar	Laptop/Handphone
4	2019	Laki-laki	3,78	Purworejo	Laptop/Handphone
5	2019	Perempuan	3,83	Medan	Laptop/Handphone
6	2019	Perempuan	3,60	Yogyakarta	Laptop/Handphone
7	2019	Laki-laki	4,00	Yogyakarta	Laptop
8	2019	Perempuan	3,94	Yogyakarta	Laptop/Handphone
...	...	...	...	...	...

Tabel 2. Contoh Data Kuesioner: Kendala Perkuliahan Daring

ID	Kendala
1	Susah bertanya saat tidak paham karena kadang terkendala sinyal
2	Kendala sinyal(Kadang sinyal hilang), kendala listrik(1.kadang mati lampu, 2. Pembayaran listrik naik sejak perkuliahan daring)
3	banyak hal yang dapat mengalihkan perhatian sehingga sulit untuk fokus pada saat kelas
4	Koneksi internet terkadang lemot, Listrik yang sering mati tapi hanya dalam beberapa detik, Orang rumah yang sering memanggil untuk keperluan ketika ada kelas, Tidak ada batas antara waktu belajar/kuliah dan istirahat, Duduk terlalu lama sehingga menyebabkan penyakit
5	Sulitnya mengerjakan tugas kelompok karena berbagai alasan seperti masalah internet dan juga niat dalam kerja secara daring.
...	...

Tabel 1 dan 2 menampilkan contoh data kuesioner yang diisi oleh responden. Tabel 1 mencakup 8 variabel yang terdiri dari informasi mengenai profil responden dan penilaian mereka terhadap perkuliahan daring. Variabel yang berhubungan dengan profil responden meliputi: tahun masuk mahasiswa, jenis kelamin, Indeks Prestasi Kumulatif (IPK), domisili, dan perangkat yang digunakan selama perkuliahan daring. Sementara itu, variabel yang terkait dengan penilaian responden meliputi: pemahaman materi, kemudahan komunikasi, dan efektivitas pelaksanaan perkuliahan daring. Semua variabel ini bersifat kuantitatif. Selain 8 variabel tersebut, terdapat satu variabel tambahan yang digunakan dalam penelitian ini, yaitu variabel kendala, yang ditampilkan pada Tabel 2. Variabel kendala ini merupakan variabel dalam bentuk data teks, di mana setiap barisnya berisi kalimat sehari-hari yang menjelaskan kendala-kendala yang dihadapi responden selama perkuliahan daring.

B. *Pre-processing* Data

Variabel kendala dalam data kuesioner, seperti yang terlihat pada Tabel 2, berbentuk teks sehingga memerlukan tahapan *pre-processing* sebelum dapat digunakan dalam proses *clustering*. Tahapan ini mencakup beberapa langkah, yaitu *case folding*, *tokenization*, *stopword removal*, dan *stemming*. Data teks menggunakan bahasa Indonesia, sehingga *stemming* dan *stopword removal* dilakukan dengan *library* Python Sastrawi. Proses ini bertujuan untuk menyederhanakan data teks dan menghilangkan elemen yang tidak relevan, seperti kata-kata umum (*stopwords*) yang tidak memberikan nilai tambah dalam analisis (Virmani Deepaliand Taneja, 2019). Setelah melalui *pre-processing*, dari total data kuesioner, sebanyak 320 data dapat digunakan, sementara 3 data lainnya dianggap tidak valid atau kosong (Mohd ariff et al., 2018).

Tabel 3. Contoh Data Kuesioner: Kendala Perkuliahan Daring

ID	Sebelum	Sesudah
1	Susah bertanya saat tidak paham karena kadang terkendala sinyal	['susah', 'tanya', 'tidak', 'paham', 'kadang', 'kendala', 'sinyal']
2	Kendala sinyal(Kadang sinyal hilang), kendala listrik(1.kadang mati lampu, 2. Pembayaran listrik naik sejak perkuliahan daring)	['kendala', 'sinyal', 'kadang', 'sinyal', 'hilang', 'kendala', 'listrik', 'kadang', 'mati', 'lampu', 'bayar', 'listrik', 'naik', 'sejak', 'kuliah', 'daring']
3	banyak hal yang dapat mengalihkan perhatian sehingga sulit untuk fokus pada saat kelas	['banyak', 'yang', 'alih', 'perhati', 'sulit', 'fokus', 'saat', 'kelas']
4	Koneksi internet terkadang lemot, Listrik yang sering mati tapi hanya dalam beberapa detik, Orang rumah yang sering memanggil untuk keperluan ketika ada kelas, Tidak ada batas antara waktu belajar/kuliah dan istirahat, Duduk terlalu lama sehingga menyebabkan penyakit	['koneksi', 'internet', 'terkadang', 'lot', 'listrik', 'sering', 'mati', 'hanya', 'beberapa', 'detik', 'orang', 'rumah', 'sering', 'panggil', 'perlu', 'ada', 'kelas', 'ada', 'batas', 'waktu', 'ajar', 'kuliah', 'istirahat', 'duduk', 'terlalu', 'lama', 'sebab', 'sakit']
5	Sulitnya mengerjakan tugas kelompok karena berbagai alasan seperti masalah internet dan juga niat dalam kerja secara daring.	['sulit', 'kerja', 'tugas', 'kelompok', 'bagai', 'alas', 'masalah', 'internet', 'juga', 'niat', 'kerja', 'daring']
...	...	...

Tabel 3 menunjukkan contoh hasil *pre-processing* yang telah dilakukan pada data kuesioner. Hasil dari tahapan *pre-processing* berupa token-token yang nantinya akan digunakan dalam tahap *clustering*. Sebagai contoh, data pada ID ke-15 awalnya mengandung tanda baca, angka, dan beberapa kata yang diawali dengan huruf kapital.

https://doi.org/10.30872/atasi.v5i1.3577

Setelah melalui *pre-processing*, kata ‘Kendala’ menjadi ‘kendala’, ‘Kadang’ menjadi ‘kadang’, dan ‘Pembayaran’ menjadi ‘pembayaran’, sedangkan angka dan tanda baca dihapus. Proses ini menghasilkan token-token seperti 'kendala', 'sinyal', 'kadang', 'sinyal', 'hilang', 'kendala', 'listrik', 'kadang', 'mati', 'lampu', 'bayar', 'listrik', 'naik', 'sejak', dan 'kuliah', yang kemudian siap digunakan untuk tahap *clustering*. Penghapusan tanda baca dan angka membantu mengurangi noise dalam data, sedangkan normalisasi kata seperti *case folding* memastikan keseragaman dalam analisis. Tahapan *pre-processing* ini sangat penting untuk meningkatkan kualitas data sebelum digunakan selanjutnya (Assiri et al., 2015).

C. *Cosine Similarity*

Penelitian ini menggunakan metode *cosine similarity* untuk mengukur tingkat kemiripan antara dua dokumen atau teks. Setiap baris data kuesioner dihitung kemiripannya dengan baris data lainnya. Nilai *cosine similarity* tersebut kemudian digunakan untuk mengelompokkan kendala yang memiliki kesamaan berdasarkan nilai kemiripan menggunakan algoritma *K-Means clustering*. Dengan teknik ini, identifikasi pola kemiripan antar kendala dapat dilakukan lebih efisien. Selain itu, pengelompokan menggunakan *K-Means* memungkinkan analisis yang lebih terfokus terhadap karakteristik dari masing-masing *cluster*, sehingga dapat memberikan wawasan lebih mendalam terkait hubungan antar kendala yang ada (Aggarwal Charu C.and Zhai, 2012).

Tabel 4. Pemetaan *Similarity* Kuesioner

	0	1	2	3	...	317	318
0	1,000000	0,000000	0,000000	0,000000	...	0,000000	0,000000
1	0,000000	1,000000	0,000000	0,258199	...	0,258199	0,000000
2	0,000000	0,000000	1,000000	0,000000	...	0,000000	0,353553
3	0,000000	0,258199	0,000000	1,000000	...	0,200000	0,158114
...	...	...	...	...	...	...	...
317	0,000000	0,258199	0,000000	0,200000	...	1,000000	0,000000
318	0,000000	0,000000	0,353553	0,158114	...	0,000000	1,000000

Tabel 4 menunjukkan contoh nilai *similarity* dari perbandingan antar kuesioner. Berdasarkan nilai tersebut, terlihat tingkat kemiripan antar kuesioner, misalnya kemiripan antara kuesioner 1 dan 3 adalah [1,3] = 0,258199 atau [3,1] = 0,258199. Tingkat kemiripan ini diukur berdasarkan token yang dihasilkan dari proses pre-processing seperti yang dijelaskan pada tabel 3. Semakin banyak token yang sama di antara dua kuesioner, semakin tinggi nilai similarity, yang mendekati nilai 1. Nilai *similarity* ini mencerminkan seberapa mirip konten dari dua kuesioner tersebut dalam hal kemunculan kata-kata atau frasa yang relevan. Semakin tinggi nilai *similarity*, semakin besar kemungkinan kedua kuesioner memiliki pola atau tema yang serupa, yang dapat membantu dalam analisis lebih lanjut.

D. *K-Means Clustering*

Algoritma *K-Means clustering* digunakan untuk mengelompokkan kendala pembelajaran daring yang dijabarkan oleh mahasiswa pada saat pengisian kuesioner. Pengelompokan atau pembuatan *cluster* dengan menggunakan nilai *similarity* seperti yang ditunjukkan pada Tabel 4 (Arora et al., 2016). Algoritma *K-Means clustering* digunakan untuk mengelompokkan kendala pembelajaran daring yang dijabarkan oleh mahasiswa saat mengisi kuesioner. Proses pengelompokan atau pembentukan *cluster* ini dilakukan berdasarkan nilai kemiripan (*similarity*) antara setiap kendala, seperti yang ditunjukkan pada Tabel 4.

Nilai *similarity* dihitung menggunakan metode *cosine similarity*, yang mampu mengukur seberapa mirip dua kendala dalam hal teks. Dengan metode ini, kendala-kendala yang memiliki karakteristik atau tema yang sama akan dikelompokkan ke dalam satu *cluster*, sehingga memudahkan analisis lebih lanjut terkait pola umum dari kendala pembelajaran daring yang dihadapi mahasiswa (Shaik et al., 2023). Penggunaan *K-Means* memungkinkan identifikasi kelompok masalah yang paling sering muncul, yang kemudian dapat digunakan sebagai dasar untuk perbaikan sistem pembelajaran daring.

Berdasarkan perhitungan menggunakan metode *Silhouette*, jumlah cluster yang paling optimal adalah 5 *cluster*, dengan nilai *Silhouette* sebesar 0,223304. Nilai K=5 dipilih setelah dilakukan pengujian pada beberapa nilai K (2–10), di mana K=5 memberikan keseimbangan terbaik antara pemisahan antar-cluster dan kemudahan interpretasi hasil. Skor 0,223304 tergolong rendah hingga moderat, namun dapat diterima mengingat data yang digunakan berbentuk teks kuesioner dengan tema yang sering tumpang tindih. Meski demikian, nilai ini tetap menunjukkan adanya struktur cluster yang dapat diinterpretasikan (Kumar Sutrakar & Mogre, 2025). Kelima cluster tersebut adalah *cluster* 0, *cluster* 1, *cluster* 2, *cluster* 3, dan *cluster* 4. *Centroid* untuk masing-masing *cluster* adalah sebagai berikut:
cluster 0 = [-1,65222558, -0,82346108],
cluster 1 = [4,21357013, -0,25687771],
cluster 2 = [1,41948226, -0,32526196],
cluster 3 = [-1,68472652, 1,42754685], dan
cluster 4 = [1,46859395, 1,87815538].

https://doi.org/10.30872/atasi.v5i1.3577

Hasil ini menunjukkan bahwa setiap *cluster* memiliki karakteristik unik yang diwakili oleh nilai *centroid*-nya. Skor *Silhouette* yang diperoleh juga mengindikasikan bahwa meskipun ada pemisahan antar-*cluster* yang jelas, kualitas *clustering* ini masih bisa ditingkatkan.

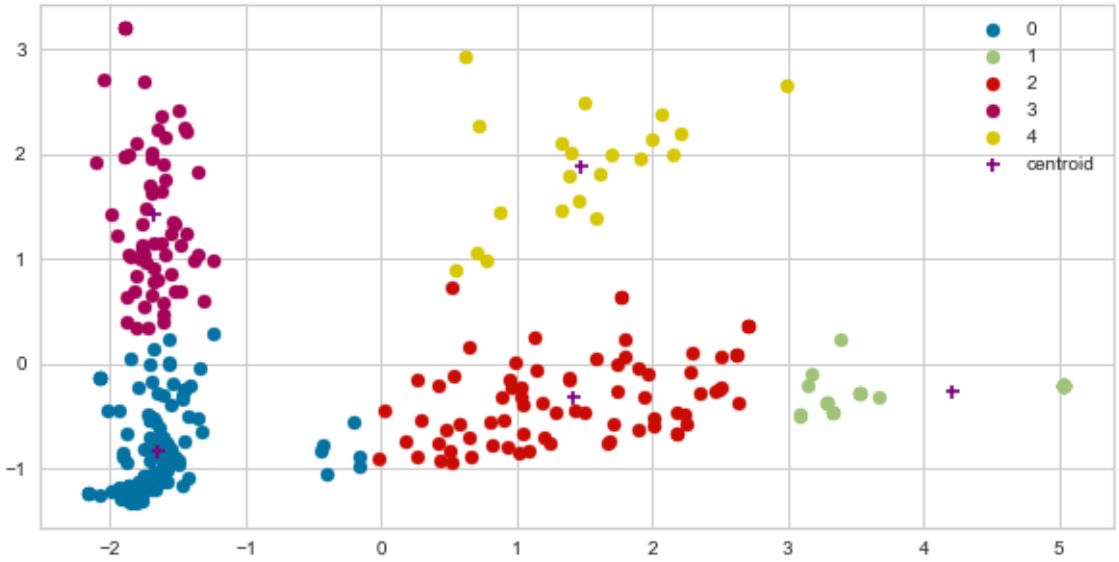
E. Hasil *Cluster* dan Pembahasan

Tabel 5 menampilkan hasil *clustering* yang mencakup kata-kata kunci yang menjadi karakteristik utama setiap *cluster*, jumlah data yang terkelompok dalam setiap *cluster*, serta persentase kecocokan masing-masing data dengan *cluster* yang relevan. Proses *clustering* ini tidak hanya membantu mengidentifikasi pola umum yang tersembunyi dalam *dataset*, tetapi juga memungkinkan analisis lebih lanjut terhadap hubungan antar data yang sebelumnya sulit dilihat secara manual (Sinaga & Yang, 2020).

Tabel 5. Hasil *Clustering*

Cluster	Nama Cluster	Kata Kunci pada Cluster	Jumlah Data	%
0	Sinyal internet tidak memadai	['sinyal', 'internet', 'jaring, 'koneksi', 'putus', 'stabil']	118	89,83
1	Lingkungan belajar tidak kondusif	['ganggu', 'panggil', 'orang', 'tamu', 'fokus', 'suasana', 'dukung', 'kondusif']	39	79,49
2	Pemahaman materi kurang maksimal, terutama untuk mata kuliah praktik	['paham', 'materi', 'ajar', 'susah', 'tangkap', 'praktek']	77	92,21
3	Sulitnya berkomunikasi dengan dosen	['sulit', 'komunikasi', 'dosen']	65	86,15
4	Sulit bekerja dalam kelompok	['sulit', 'tugas', 'teman', 'kelompok']	21	76,19

Dari Tabel 5 dapat disimpulkan bahwa setiap *cluster* berisi data yang memiliki keterkaitan erat. Pengelompokan ini menunjukkan bahwa data dalam satu *cluster* cenderung memiliki pola atau karakteristik yang serupa, yang secara signifikan memperkaya pemahaman tentang dinamika yang ada dalam dataset. Dari hasil *clustering* yang telah terbentuk, ditemukan bahwa terdapat beberapa data yang tidak sesuai dengan kata kunci, sehingga data tersebut seharusnya tidak berada di dalam *cluster* tersebut. Tingkat kecocokan dalam setiap *cluster* adalah sebagai berikut: 89,83% untuk *cluster* 0, 79,49% untuk *cluster* 1, 92,21% untuk *cluster* 2, 86,15% untuk *cluster* 3, dan 76,19% untuk *cluster* 4.



Gambar 3. Distribusi kendala perkuliahan daring berdasarkan pemetaan *cluster* yang dihasilkan dari metode *K-Means* dengan 5 *cluster*

Distribusi atau sebaran hasil *cluster* yang telah terbentuk dan *centroid*-nya dapat dilihat pada Gambar 3. Dari gambar tersebut dapat terlihat sebaran masing-masing data dengan *cluster* yang telah terbentuk. Berdasarkan analisis data dan identifikasi kata kunci pada setiap *cluster*, *cluster* 0 dinamai "sinyal internet tidak memadai". *Cluster* ini menampung data terbanyak, yaitu sebanyak 118 data, dengan tingkat akurasi mencapai 89,83%. Temuan ini menunjukkan bahwa kendala utama yang dialami mahasiswa selama perkuliahan daring adalah kesulitan memperoleh sinyal internet yang memadai, yang secara signifikan menghambat kelancaran mereka dalam mengikuti perkuliahan. Faktor geografis dan keterbatasan akses terhadap infrastruktur internet di beberapa wilayah menjadi tantangan yang mendasar.

https://doi.org/10.30872/atasi.v5i1.3577

Cluster yang memiliki *centroid* terdekat dengan *cluster* 0 adalah *cluster* 2, *cluster* 3, dan *cluster* 4. Pada *cluster* 2, yang diberi nama "pemahaman materi kurang maksimal, terutama untuk mata kuliah praktik", terdapat sebanyak 77 data dengan tingkat akurasi sebesar 92,21%. Data dalam *cluster* 2 menunjukkan bahwa mahasiswa menghadapi kendala dalam memahami materi yang disampaikan, terutama ketika mata kuliah tersebut berbasis praktik. Hal ini menandakan bahwa mahasiswa merasa kesulitan dalam mengikuti kegiatan praktik yang biasanya memerlukan bimbingan langsung atau interaksi yang intensif dengan dosen atau asisten dosen. Akibatnya, mahasiswa terpaksa melaksanakan kegiatan praktik secara mandiri, yang dapat mengurangi efektivitas pembelajaran dan menimbulkan kesenjangan pemahaman.

Data yang tergolong dalam cluster 3 berjumlah 65, dengan tingkat akurasi mencapai 86,15%, sedangkan *cluster* 4 terdiri dari 21 data dengan tingkat akurasi 76,19%. *Cluster* 3 dinamai "sulitnya berkomunikasi dengan dosen", mencerminkan kesulitan mahasiswa dalam berinteraksi dan memperoleh bimbingan dari dosen selama perkuliahan daring. Sementara itu, cluster 4 diberi nama "sulit bekerja dalam kelompok", menggambarkan tantangan mahasiswa dalam berkolaborasi dengan rekan saat menyelesaikan tugas kelompok secara daring. Kedua cluster ini berakar pada permasalahan serupa, yaitu tantangan dalam interaksi yang terjadi antara mahasiswa, baik dengan dosen maupun sesama mahasiswa. Hambatan-hambatan tersebut dapat disebabkan oleh beberapa faktor, seperti keterbatasan komunikasi non-verbal, kendala teknis dalam platform daring, serta kurangnya keakraban antar anggota kelompok atau antara mahasiswa dengan dosen.

Cluster 1 diberi nama "lingkungan belajar tidak kondusif," dengan jumlah data sebanyak 39 dan tingkat akurasi sebesar 79,49%. *Cluster* ini menunjukkan bahwa mahasiswa menghadapi lingkungan belajar yang kurang mendukung, sehingga menghambat efektivitas mereka dalam mengikuti perkuliahan daring. Berdasarkan analisis data pada cluster ini, lingkungan belajar yang tidak kondusif disebabkan oleh berbagai faktor, termasuk suasana atau lingkungan sekitar yang menimbulkan gangguan, baik yang disengaja maupun tidak. Gangguan tersebut bisa berupa kebisingan dari anggota keluarga atau tetangga, kondisi rumah yang ramai, hingga keterbatasan fasilitas pendukung belajar seperti ruang pribadi yang tenang. Kondisi ini menimbulkan dampak negatif pada konsentrasi dan produktivitas mahasiswa, membuat proses pembelajaran menjadi kurang optimal.

Hasil *clustering* dalam penelitian ini mengungkapkan beberapa pola karakteristik yang signifikan di antara data yang dianalisis, dengan masing-masing *cluster* menunjukkan tema-tema spesifik terkait kendala yang dihadapi mahasiswa dalam perkuliahan daring. Pengelompokan data ini mempermudah identifikasi masalah-masalah utama yang dihadapi, seperti sinyal internet yang tidak memadai, lingkungan belajar yang kurang kondusif, kesulitan dalam memahami materi praktik, serta hambatan dalam komunikasi dengan dosen dan kolaborasi dalam tugas kelompok. Pengelompokan ini juga memungkinkan kita melihat hubungan antara karakteristik data dalam satu *cluster*, memperkaya pemahaman tentang faktor-faktor yang memengaruhi pengalaman belajar mahasiswa.

Klasifikasi ini menunjukkan bahwa kesulitan dalam memperoleh sinyal internet yang stabil menjadi masalah utama yang dirasakan mahasiswa, terutama di wilayah dengan akses internet yang terbatas. Tantangan teknis ini diikuti oleh kendala dalam pemahaman materi berbasis praktik, di mana mahasiswa merasa terbatas dalam proses belajar karena keterbatasan bimbingan langsung dari dosen. Pola-pola yang ditemukan ini mengindikasikan adanya kesenjangan dalam pengalaman belajar daring yang mungkin berpengaruh terhadap efektivitas pembelajaran.

Analisis lebih lanjut juga menunjukkan bahwa tantangan dalam komunikasi dengan dosen dan kolaborasi kelompok dalam perkuliahan daring menjadi faktor yang menghambat pengalaman akademik mahasiswa. *Cluster* yang berkaitan dengan komunikasi ini mencerminkan permasalahan yang dihadapi dalam berinteraksi secara efektif, baik karena kendala teknis maupun sosial, seperti kurangnya dukungan lingkungan belajar. Dengan memahami pola-pola ini, pihak terkait dapat menyusun strategi yang lebih terfokus untuk mengatasi masalah-masalah spesifik dalam pembelajaran daring, guna meningkatkan kualitas dan kenyamanan belajar bagi mahasiswa.

5. KESIMPULAN

Penelitian ini menunjukkan variasi tingkat kesiapan mahasiswa dalam perkuliahan daring, yang seringkali dipengaruhi oleh hambatan seperti koneksi internet, perangkat, lingkungan belajar, dan dukungan keluarga. Hambatan-hambatan ini dikelompokkan menjadi beberapa kategori untuk memudahkan analisis dan penemuan solusi yang tepat. Tantangan terbesar dalam penelitian ini adalah jumlah data yang besar yang membutuhkan analisis menyeluruh. Dengan mengelompokkan data berdasarkan karakteristik yang serupa, penelitian ini dapat membantu universitas dalam memahami dan mengatasi kendala utama yang dihadapi mahasiswa.

Untuk analisis, metode *K-Means Clustering* digunakan untuk mengelompokkan data survei mahasiswa berdasarkan kesamaan dalam kendala yang mereka alami. Data yang melalui tahap *pre-processing* diukur tingkat kemiripannya menggunakan *cosine similarity* untuk kemudian dikelompokkan dengan *K-Means*. Hasil dari clustering ini mengidentifikasi beberapa kategori kendala utama seperti masalah koneksi internet, pemahaman materi praktik, dan kesulitan komunikasi dan kolaborasi. Dengan pendekatan ini, universitas dapat lebih mudah memahami faktor-faktor yang menghambat pengalaman belajar mahasiswa dalam pembelajaran daring.

https://doi.org/10.30872/atasi.v5i1.3577

Hasil analisis ini menunjukkan bahwa kendala terbesar adalah koneksi internet yang tidak stabil, yang diikuti oleh hambatan dalam memahami materi praktik, dan tantangan dalam komunikasi dan kolaborasi daring. Temuan ini memberikan wawasan penting bagi universitas untuk meningkatkan efektivitas pembelajaran daring, seperti dengan menyusun strategi yang lebih adaptif terhadap kendala teknis dan sosial yang dihadapi mahasiswa. Dengan pemahaman ini, universitas dapat merencanakan langkah-langkah yang lebih efektif dalam mendukung proses belajar-mengajar daring, guna meningkatkan kualitas pendidikan bagi mahasiswa.

6. UCAPAN TERIMA KASIH

Artikel jurnal ini merupakan salah satu hasil dari penelitian yang didanai oleh Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Atma Jaya Yogyakarta. Penelitian ini bertujuan untuk memberikan kontribusi signifikan dalam pengembangan ilmu pengetahuan dan pemecahan masalah yang relevan di masyarakat.

7. DAFTAR PUSTAKA

Abbas, W., Ahmed, M., Khalid, R., & Yasmeen, T. (2017). Analyzing the factors that can limit the acceptability to introduce new specializations in higher education institutions: A case study of higher education institutions of Southern Punjab, Pakistan. *International Journal of Educational Management*, *31*, 530–539. https://doi.org/10.1108/IJEM-06-2016-0139

Adedoyin, O. B., & Soykan, E. (2023). Covid-19 pandemic and online learning: the challenges and opportunities. *Interactive Learning Environments*, *31*(2), 863–875. https://doi.org/10.1080/10494820.2020.1813180

Adnan, M. (2020). Online learning amid the COVID-19 pandemic: Students perspectives. *Journal of Pedagogical Sociology and Psychology*, *1*(2), 45–51. https://doi.org/10.33902/jpsp.2020261309

Aggarwal Charu C. and Zhai, C. (2012). A Survey of Text Clustering Algorithms. In C. Aggarwal Charu C. and Zhai (Ed.), *Mining Text Data* (pp. 77–128). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_4

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science*, *78*, 507–512. https://doi.org/https://doi.org/10.1016/j.procs.2016.02.095

Assiri, A., Emam, A., & Aldossari, H. (2015). Arabic Sentiment Analysis: A Survey. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 6, Issue 12). www.ijacsa.thesai.org

Bustami, Nurdin, Taufiq, Fajriana, & Khatami, M. (2023). Application of the K-Means Clustering Algorithm to Determine Student Academic Ability Levels during Covid-19. *Journal of Advanced Zoology*, *44*(3), 1215–1226. https://jazindia.com

Dharma, R. (2019, May 6). *The Math Behind K-Means Clustering*. Medium. https://medium.com/@draj0718/the-math-behind-k-means-clustering-4aa85532085e

Dhawan, S. (2020). Online Learning: A Panacea in the Time of COVID-19 Crisis. *Journal of Educational Technology Systems*, *49*(1), 5–22. https://doi.org/10.1177/0047239520934018

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. In *Journal of Emerging Technologies in Web Intelligence* (Vol. 1, Issue 1, pp. 60–76). Academy Publisher. https://doi.org/10.4304/jetwi.1.1.60-76

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/J.INS.2022.11.139

Jianqiang, Z., & Xiaolin, G. (2017). Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access*, *5*, 2870–2879. https://doi.org/10.1109/ACCESS.2017.2672677

Kumar, N., Yadav, S. K., & Yadav, D. S. (2020). Similarity Measure Approaches Applied in Text Document Clustering for Information Retrieval. *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 88–92. https://doi.org/10.1109/PDGC50313.2020.9315851

Kumar Yadav, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification Saurabh Pal. In *World of Computer Science and Information Technology Journal (WCSIT)* (Vol. 2, Issue 2).

Mohd ariff, N., Abu Bakar, M. A., & Rahmad, M. I. (2018). Comparative Study of Document Clustering Algorithms. *International Journal of Engineering and Technology(UAE)*, *7*, 246–251. https://doi.org/10.14419/ijet.v7i4.11.20816

Münz, G., Li, S., & Carle, G. (2007). Traffic Anomaly Detection Using K-Means Clustering. *International Workshop on Traffic Analysis and Classification*. https://api.semanticscholar.org/CorpusID:2303041

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*, 100003. https://doi.org/https://doi.org/10.1016/j.nlp.2022.100003

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

https://doi.org/10.30872/atasi.v5i1.3577

Singh, I., Sabitha, A. S., & Bansal, A. (2016). Student performance analysis using clustering algorithm. *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, 294–299. https://doi.org/10.1109/CONFLUENCE.2016.7508131

Kumar Sutrakar, V., & Mogre, N. (2025). An Improved Deep Learning Model for Word Embeddings Based Clustering for Large Text Datasets. *Machine Learning Research*, *10*(1), 32. https://doi.org/10.11648/j.mlr.20251001.14

Virmani Deepali and Taneja, S. (2019). A Text Preprocessing Approach for Efficacious Information Retrieval. In M. C. and M. K. K. and T. S. and S. P. K. Panigrahi Bijaya Ketan and Trivedi (Ed.), *Smart Innovations in Communication and Computational Sciences* (pp. 13–22). Springer Singapore.

https://doi.org/10.30872/atasi.v5i1.3577