

INTEGRASE K-MEANS, LATENT DIRICHLET ALLOCATION, DAN ANALISIS SENTIMEN UNTUK ANALISIS MULTIDIMENSI UMPAN BALIK MAHASISWA

Paulus Ola Laga¹⁾, Abdul Aziz²⁾, Syahminan³⁾

Universitas PGRI Kanjuruhan Malang^{1,2,3)}
ragenpaull@gmail.com

Abstrak

Penelitian ini bertujuan mengelompokkan komentar mahasiswa terhadap dosen pada Sistem Informasi Akademik (SIKAD) Universitas PGRI Kanjuruhan Malang menggunakan algoritma K-Means. Kami menganalisis 33.229 komentar melalui tahap pra-pemrosesan teks (*normalization, tokenization, stopword removal, dan stemming*). Selanjutnya, komentar kami transformasikan menjadi representasi numerik dengan metode Term Frequency-Inverse Document Frequency (TF-IDF). Kami menentukan jumlah kluster optimal melalui metode Elbow yang menghasilkan tiga kluster. Evaluasi kualitas kluster menggunakan Silhouette Score memperlihatkan nilai relatif rendah (0,084). Meski demikian, analisis lanjutan dengan Latent Dirichlet Allocation (LDA) dan analisis sentimen berbasis kosakata mengungkap karakteristik unik tiap kluster: (1) didominasi apresiasi umum (baik, ramah), (2) berfokus pada aspek pembelajaran (materi jelas, metode mengajar), dan (3) berisi komentar personal dan kritik spesifik (tugas banyak, nilai lambat). Temuan ini membuktikan bahwa klusterisasi komentar yang diperkuat analisis topik dan sentimen dapat menyediakan insight strategis bagi universitas guna mendukung pengambilan keputusan akademik (seperti pengembangan program dosen) serta meningkatkan mutu pengajaran secara terfokus.

Kata Kunci : K-Means clustering; student feedback analysis; text preprocessing; Latent Dirichlet Allocation
Lexicon-based sentiment analysis

Abstract

This study aims to cluster student feedback on lecturers in the Academic Information System (SIKAD) of Universitas PGRI Kanjuruhan Malang using the K-Means algorithm. We analyzed 33,229 comments through text preprocessing stages (*normalization, tokenization, stopword removal, and stemming*). Subsequently, we transformed the comments into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method. We determined the optimal cluster number via the Elbow method, yielding three clusters. Evaluation of cluster quality using the Silhouette Score revealed a relatively low value (0.084). Nevertheless, further analysis with Latent Dirichlet Allocation (LDA) and lexicon-based sentiment analysis uncovered unique characteristics for each cluster: (1) dominated by general appreciation (good, friendly), (2) focused on learning aspects (clear material, teaching methods), and (3) containing personal comments and specific criticisms (excessive assignments, slow grading). These findings demonstrate that comment clustering complemented by topic modeling and sentiment analysis can provide strategic insights for the university to support academic decision-making (e.g., lecturer development programs) and enhance teaching quality in a targeted manner.

Keywords : K-Means clustering; student feedback analysis; text preprocessing; Latent Dirichlet Allocation
Lexicon-based sentiment analysis

1. PENDAHULUAN

Pendidikan tinggi di era digital berperan kritis dalam menyiapkan lulusan yang kompeten dan berkarakter. Evaluasi proses pembelajaran menjadi komponen vital untuk menjamin kualitas pendidikan tersebut. Perkembangan teknologi informasi mendorong adopsi sistem digital seperti Sistem Informasi Akademik (SIKAD), yang tidak hanya mengelola data akademik tetapi juga memfasilitasi umpan balik mahasiswa terhadap kinerja dosen. Komentar mahasiswa ini merupakan sumber data strategis untuk meningkatkan mutu pengajaran.

Universitas PGRI Kanjuruhan Malang (Unikama) mengandalkan SIAKAD sebagai platform utama pengumpulan evaluasi dosen oleh mahasiswa. Namun, data komentar yang terkumpul bersifat unstructured (tidak terstruktur), heterogen, dan dalam volume besar. Analisis manual terhadap data semacam ini terbukti tidak efisien: memakan waktu, rentan kesalahan, dan kurang skalabel. Tantangan ini diperparah oleh tingkat partisipasi mahasiswa yang belum optimal (40%-60%), sehingga berpotensi mengurangi representativitas wawasan yang dihasilkan.

Pendekatan data mining seperti algoritma K-Means Clustering diusulkan sebagai solusi potensial untuk mengatasi kompleksitas analisis data tekstual. Algoritma ini mampu mengelompokkan komentar secara otomatis berdasarkan kemiripan karakteristik (misalnya: sentimen positif, negatif, atau netral). Meskipun K-Means telah teruji dalam studi serupa (e.g., analisis sentimen layanan akademik) (Mashudi et al., 2023) (Zein, Farizy, & Suharyanto, 2022) (Akbar & Salam, 2024) (Muhariya, 2022), implementasinya secara spesifik untuk klasterisasi komentar evaluatif dosen pada SIAKAD konteks lokal masih belum banyak dieksplorasi. Fokus inilah yang membedakan penelitian ini dari pendahulunya.

Penelitian ini menawarkan orisinalitas melalui tiga aspek: (1) Objek spesifik: komentar mahasiswa tentang kinerja dosen di SIAKAD Unikama, (2) Tujuan aplikatif: meningkatkan kualitas pembelajaran melalui wawasan terstruktur, dan (3) Relevansi institusional: desain solusi sesuai kebutuhan lokal perguruan tinggi. Integrasi langsung dengan sistem akademik kampus juga menjadi pembeda dari penelitian umum tentang analisis sentimen pendidikan.

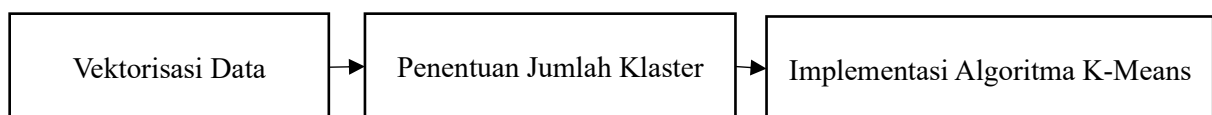
Secara khusus, penelitian ini bertujuan mengelompokkan komentar mahasiswa terhadap dosen pada SIAKAD Unikama menggunakan algoritma K-Means guna mengidentifikasi pola umpan balik yang terstruktur. Dampak yang diharapkan mencakup: peningkatan efisiensi analisis komentar, dukungan bagi dosen dalam mengevaluasi metode pengajaran, serta penyediaan dasar data untuk pengambilan keputusan strategis (e.g., pengembangan kurikulum, pelatihan dosen, dan kebijakan peningkatan mutu pembelajaran).

2. METODE

2.1 Sumber Data dan Karakteristik

Data penelitian berupa 33.229 komentar teks tidak terstruktur diambil secara retrospektif dari modul evaluasi dosen pada Sistem Informasi Akademik (SIAKAD) Universitas PGRI Kanjuruhan Malang. Data mencakup periode akademik tahun 2021 hingga 2024 dan terdiri atas tanggapan anonim mahasiswa dalam Bahasa Indonesia kolokial (contoh: 'Dosenya ramah dan jelas mengajarnya', 'Tugas terlalu banyak'). Proses ekstraksi data telah memperoleh persetujuan institusi dengan menjamin anonimitas responden melalui penghapusan identitas pribadi (nama mahasiswa/dosen) dan metadata sensitif.

2.2 Klasterisasi



Gambar 1.1 Tahapan Klasterisasi

Teks yang telah diproses ditransformasikan menjadi representasi numerik menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) dengan pembatasan dimensi fitur sebanyak 1.000 term berperingkat tertinggi berdasarkan skor IDF untuk optimasi komputasi. Penentuan jumlah klaster optimal (k) dilakukan melalui Elbow Method dengan menganalisis tren penurunan *Within-Cluster Sum of Squares* (WCSS) pada rentang nilai $k=2$ hingga 10. Berdasarkan identifikasi titik tekuk grafik, nilai k optimal ditetapkan. Selanjutnya, algoritma

K-Means diimplementasikan menggunakan library Scikit-learn dengan konfigurasi parameter **random_state=42** (jaminan replikabilitas), **n_init=10** (jumlah inisialisasi centroid), dan iterasi maksimum 300. Evaluasi objektif kualitas partisi kluster dilakukan menggunakan metrik Silhouette Score.

Pasca-proses klasterisasi, dilakukan analisis lanjutan yang mencakup tiga langkah utama. Pertama, topic modeling menggunakan *Latent Dirichlet Allocation* (LDA) diterapkan untuk mengekstraksi topik utama dalam setiap kluster. Selanjutnya, analisis sentimen berbasis leksikon (kosakata) digunakan guna mengidentifikasi persepsi emosional teks, yaitu sentimen positif, netral, dan negatif. Untuk mempermudah interpretasi karakteristik unik tiap kluster secara inovatif, dilakukan visualisasi data menggunakan berbagai teknik: PCA 2D (*Principal Component Analysis*) untuk reduksi dimensi, Word Cloud untuk menampilkan kata kunci dominan, Bar Chart untuk perbandingan kuantitatif, serta Heatmap untuk menggambarkan hubungan antar-variabel.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari modul evaluasi dosen ada Sistem Informasi Akademik (SIKAD) Universitas PGRI Kanjuruhan Malang. Komentar dikumpulkan dari tahun akademik 2021 hingga 2024, dengan total 33.229 entri komentar teks. Data dikumpulkan dalam format CSV, yang terdiri dari dua kolom utama: tahun akademik dan teks komentar. Proses pengambilan data dilakukan melalui izin dari pihak pengelola sistem akademik dan dilakukan secara anonim untuk menjaga kerahasiaan identitas mahasiswa.

Tabel 3. 1 Data awal komentar mahasiswa

No	Tahun	Komentar Mahasiswa
1	20212	Terimakasih untuk bu henni yang sudah membimbing saya dan teman-teman dalam mata kuliah ini, meski disayangkan untuk perkuliahan lebih banyak daring. semoga kedepannya saya dapat di bimbing kembali oleh bu henni dan dapat melakukan perkuliahan secara luring.
2	20212	Sangat baik dalam melakukan dan mengatur proses pembelajaran
3	20212	Sangat baik dalam proses pembelajaran
4	20212	Sangat baik dalam proses pembelajaran
5	20212	Sangat baik dalam proses pembelajaran
...	...	...
33225	20241	Sangat baik
33226	20241	Sangat baik
33227	20241	Sangat baik
33228	20241	Sangat baik
33229	20241	baik sekli

3.2 Pra-pemrosesan Data

Sebelum dilakukan analisis kluster, data teks komentar perlu dipersiapkan melalui serangkaian tahap pra-pemrosesan teks untuk memastikan kebersihan dan konsistensi format. Tahapan yang dilakukan adalah:

- Pembersihan teks: Menghapus angka, simbol, tanda baca, serta mengubah huruf menjadi kecil (case foding).
- Tokenisasi: Memecah kalimat menjadi kata-kata individual (token).
- Stopword removal: Menghapus kata-kata umum yang tidak memiliki makna signifikan dalam analisis (seperti “yang”, “dan”, “dalam”).

- d. Stemming: Mengubah kata kebentuk dasarnya (misal: “mengajar”, “pengajaran” menjadi “ajar”) menggunakan library sastrawi yang cocok untuk bahasa indonesia.

Tabel 3. 2 Data Hasil Pra-pemrosesan

No	Tahun Ajaran	Komentar Asli	Komentar Bersih	Tokenisasi	Hasil Stemming
1	20212	Terimakasih untuk bu henni yang sudah membimbing saya dan teman-teman...	terimakasih untuk bu henni yang sudah membimbing saya dan teman teman	['terimakasih', 'untuk', 'bu', 'henni', ...]	terimakasih untuk bu henni ajar saya teman teman
2	20212	Sangat baik dalam melakukan dan mengatur proses pembelajaran	sangat baik dalam melakukan dan mengatur proses pembelajaran	['sangat', 'baik', 'dalam', 'mengatur', 'proses', 'melakukan', ...]	sangat baik atur proses ajar
3	20212	Sangat baik dalam proses pembelajaran	sangat baik dalam proses pembelajaran	['sangat', 'baik', 'dalam', 'proses', 'pembelajaran']	sangat baik proses ajar
4	20212	Sangat baik dalam proses pembelajaran	sangat baik dalam proses pembelajaran	['sangat', 'baik', 'dalam', 'proses', 'pembelajaran']	sangat baik proses ajar
5	20212	Sangat baik dalam proses pembelajaran	sangat baik dalam proses pembelajaran	['sangat', 'baik', 'dalam', 'proses', 'pembelajaran']	sangat baik proses ajar
..					
33225	20241	Sangat baik	sangat baik	['sangat', 'baik']	sangat baik
33226	20241	Sangat baik	sangat baik	['sangat', 'baik']	sangat baik
33227	20241	Sangat baik	sangat baik	['sangat', 'baik']	sangat baik
33228	20241	Sangat baik	sangat baik	['sangat', 'baik']	sangat baik
33229	20241	baik sekli	baik sekli	['baik', 'sekli']	baik sekali

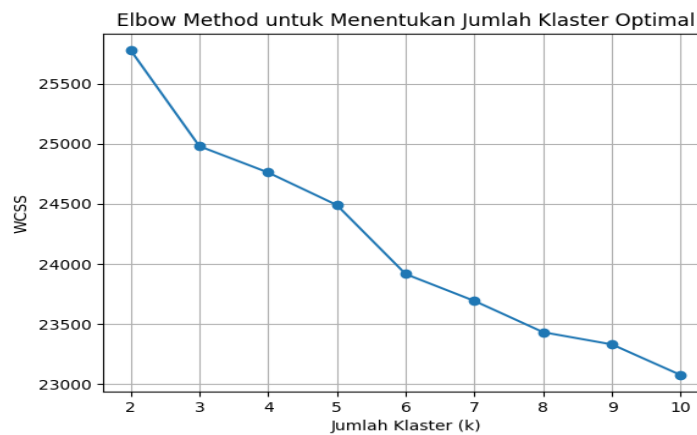
3.3 Hasil Klasterisasi

Setelah dilakukan proses TF-IDF dan penerapan algoritma K-Means, diperoleh hasil bahwa jumlah klaster optimal adalah tiga ($k=3$) berdasarkan Elbow Method. Setiap komentar mahasiswa diberi lebel klaster (0, 1, atau 2) sesuai dengan kesamaan vektornya.

Distribusi komentar per klaster:

- Klaster 0: 24.230 komentar (72,9%)
- Klaster 1: 5.636 komentar (17%)
- Klaster 2 3.363 komentar (10,1%)

Ketidakseimbangan distribusi ini wajar terjadi karena K-Means berfokus pada kemiripan fitur, bukan jumlah data yang seimbang.

**Gambar 3. 1 Grafik Elbow Method**

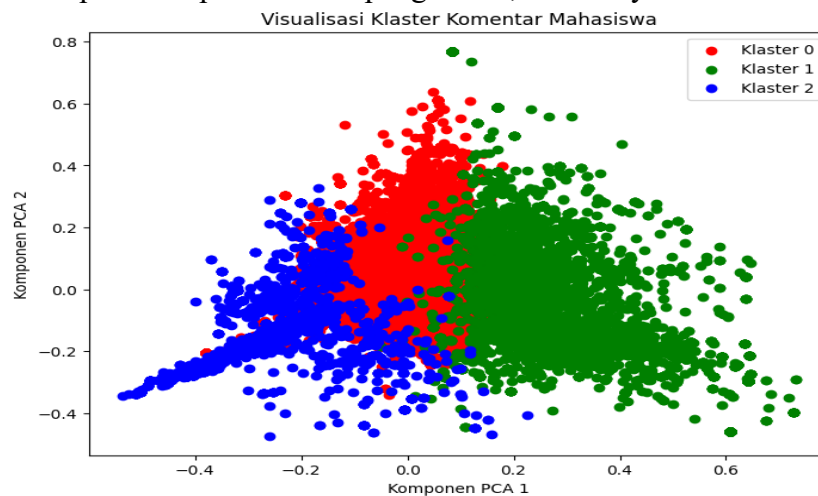
3.4 Evaluasi Kualitas Kluster

Evaluasi dilakukan menggunakan Silhouette Score, dengan hasil: 0,084

Nilai ini menunjukkan bahwa kualitas pemisahan antar kluster cukup baik meskipun tidak terlalu tinggi, mengingat:

- Komentar bersifat pendek dan umum
- Banyak komentar memiliki sentimen campuran (multitopik)

Visualisasi menggunakan PCA 2D menunjukkan bahwa ketiga kluster memiliki area dominasi yang berbeda, meskipun terdapat zona tumpang tindih, khususnya antara kluster 0 dan 1.

**Gambar 3. 2 Visualisasi PCA 2D**

3.5 Analisis Topik Tiap Kluster

Analisis topik dengan Latent Dirichle Allocation (LDA) menunjukkan:

- Kluster 0 (umum dan apresiatif): Dominan kata seperti *baik*, *terima kasih*, *mata kuliah*
- Kluster 1 (pembelajaran): Muncul kata *mudah*, *materi*, *jelas*, *paham*
- Kluster 2 (personal dan doa): Terdapat kata seperti *semoga*, *bu*, *bimbingan*, *ilmu*.

3.6 Analisis Sentimen Tiap Kluster

Sentimen komentar diidentifikasi menggunakan pendekatan berbasis kosakata:

- Kluster 0: Dominasi komentar positif (46%) dan netral (44%)
- Kluster 1: Sentimen positif mendominasi (61%), terkait penyampaian materi
- Kluster 2: Mayoritas netral (78%) berisi ucapan terima kasih dan harapan.

3.7 Visualisasi Word Cloud, Bar Chart, dan Heatmap

Untuk memperkuat hasil pemahaman hasil, digunakan beberapa metode visualisasi:

Integrasi K-Means, Latent Dirichlet Allocation, dan Analisis Sentimen Untuk Analisis Multidimensi Umpan Balik Mahasiswa

- WordCloud Klaster 0



Gambar 3. 3 Visualisasi Word Cloud Kluster 0

- WordCloud Klaster 1



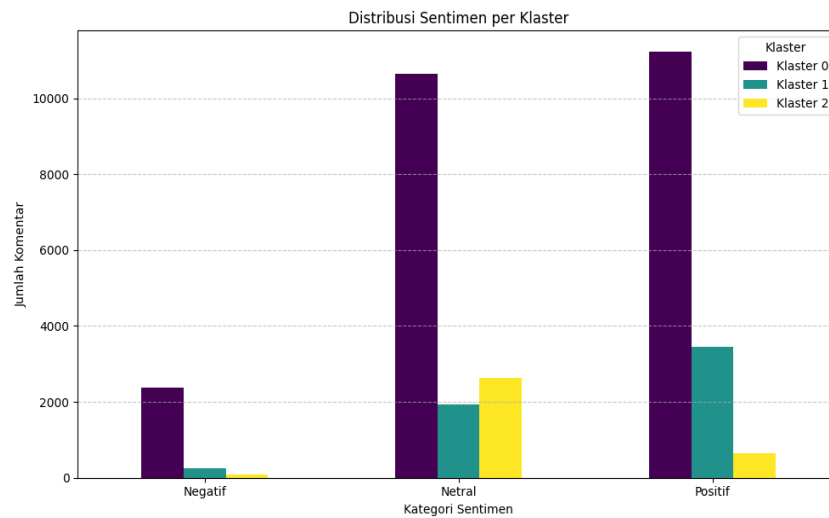
Gambar 3. 4 Visualisasi Word Cloud Cluster 1

- WordCloud Klaster 2



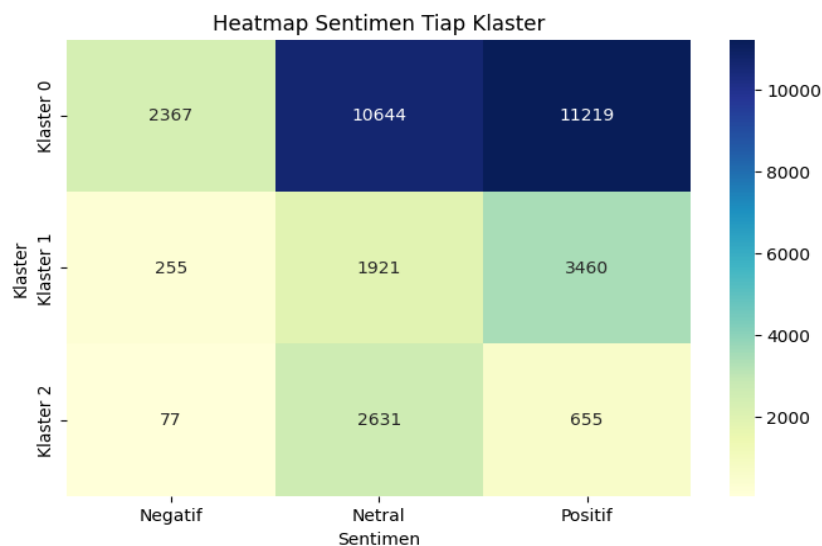
Gambar 3. 5 Visualisasi Word Cloud Cluster 2

- b. Bar Chart: Distribusi jumlah komentar berdasarkan sentimen tiap klaster



Gambar 3. 6 Visualisasi Bar Chart

- c. Heatmap: Memperlihatkan intensitas dan distribusi sentimen secara komparatif antar klaster



Gambar 3. 7 Visualisasi Heatmap

3.8 Temuan Penting

- Sebanyak 8.265 komentar (24,9%) berada pada zona ambang klaster, menunjukkan adanya komentar multitopic.
- K-Means tetap mampu menghasilkan klaster yang bermakna meskipun data teks komentar sangat singkat dan ambigu.
- Klasterisasi mampu mengungkap sentiment tersembunyi dan topik yang tidak langsung terlihat oleh dosen secara manual.

4. KESIMPULAN

Kesimpulan

Penelitian ini berhasil menerapkan algoritma K-Means untuk mengelompokkan 33.229 komentar mahasiswa terhadap dosen dari Sistem Informasi Akademik (SIKAD) Universitas

PGRI Kanjuruhan Malang. Dengan tahapan prapemrosesan teks transformasi TF-IDF, dan evaluasi melalui Elbow Method serta Silhouette Score, diperoleh hasil sebagai berikut:

- a. Jumlah kluster optimal adalah tiga dengan karakteristik berbeda:
 1. Kluster 0: Komentar umum dan apresiatif (positif/netral)
 2. Kluster 1: Komentar terkait kualitas pengajaran dan materi (positif dominan)
 3. Kluster 2: Komentar personal seperti ucapan terimakasih (netral dominan)
- b. Evaluasi dengan silhouette score sebesar 0,084 menunjukkan pemisahan kluster yang cukup baik meski terdapat tumpang tindih dalam tiap kluster.
- c. Visualisasi PCA, Word Cloud, Bar Chart, Heatmap mendukung interpretasi hasil secara lebih komprehensif dan komunikatif.

Saran : Beberapa saran yang dapat diberikan untuk penerapan lebih lanjut:

institusi: a). Integrasikan proses klusterisasi ke dalam system evaluasi dosen secara otomatis di SIAKAD. B). Kembangkan dashboard visualisasi hasil evaluasi berbasis kluster dan sentiment. Untuk dosen: a). Gunakan hasil klusterisasi sebagai cerminan persepsi mahasiswa untuk memperbaiki metode pengajaran. B). Perhatikan komentar netral atau multitopik sebagai bahan diskusi lanjut di kelas. Untuk penelitian selanjutnya: a). Gunakan algoritma lanjutan seperti fuzzy clustering atau deep learning untuk menangani komentar multitopik yang ambigu. b).Sertakan atribut seperti program studi atau mata kuliah untuk segmentasi yang lebih tajam.

REFERENSI

- [1] A. Riza, Arif, and H. S. Pratiwi, “Implementasi Metode K-Means Clustering dan Algoritma Cosine Similarity pada Repository Digital Jurusan Informatika,” *Jurnal Aplikasi dan Riset Informatika*, vol. 2, no. 2, pp. 38–51, 2024.
- [2] Bustami, Nurdin, Taufiq, Fajriana, and M. Khatami, “Application of the K-Means Clustering Algorithm to Determine Student Academic Ability Levels during Covid-19,” *Journal Of Advanced Zoology*, vol. 44, no. 3, pp. 1215–1226, Nov. 2023.
- [3] I. Made, D. E. Ratnawati, and Issa Arwani, “Analisis Sentimen Feed Back Mahasiswa terhadap Dosen Prodi Teknologi Informasi Departemen Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 6, pp. 2758–2766, 2023, Accessed: Sep. 02, 2025.
- [4] Ismi Ana Sulasiyah, “Analisis clustering e-learning readiness di pulau jawa menggunakan k-means dan principal component analysis (pca) dengan visualisasi gi,” *Uinjkt.ac.id*, 2022.
- [5] Mashudi, I. A., Arief, S. N., E.I., D. S., Fatmawati, T., Hani’ah, M., & Alfarid, I. T. (2023). “Klusterisasi Jawaban Uraian Mahasiswa Menggunakan TF-IDF dan K-Means untuk Membantu Koreksi Ujian,” *Jurnal Media Informatika Budidarma*, 7(4), 2159.
- [6] Muhajir, R. Akbar, and A. Salam, “Analisis Kepuasan Pengguna Sistem Informasi Akademik (SIAKAD) dengan Metode End User Computing Satisfaction (EUCS) di Institut Seni Budaya Indonesia Aceh,” *Jurnal Sistem Komputer (SISKOM)*, vol. 4, no. 1, pp. 45–59, Feb. 2024,
- [7] Muhariya, A. (2022). “Pengelompokan Komentar Pada Media Sosial Instagram Menggunakan Metode K-Means Clustering Untuk Identifikasi Awal Cyberbullying,” *Universitas Islam Indonesia*, 1–71.
- [8] N. Hidayah and D. Dodiman, “Implementasi algoritma multinomial naïve bayes, TF-IDF dan confusion matrix dalam pengklasifikasian saran monitoring dan evaluasi mahasiswa terhadap dosen teknik informatika Universitas Dayanu Ikhsanuddin,” *Google.com*, 2024.
- [9] Sri Wulan Utami Vitandy, A. A. Supianto, and Fitra Abdurrachman Bachtiar, “Analisis Sentimen Evaluasi Kinerja Dosen menggunakan Term Frequency-Inverse Document

- Frequency dan Naive Bayes Classifier,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 6, pp. 6080–6088, 2019, Accessed: Sep. 02, 2025.
- [10] T. Hidayat, Rinda Cahyana, and Indri Tri Julianto, “Analisis Sentimen Layanan Sistem Informasi Akademik Mahasiswa Menggunakan Algoritma Naive Bayes,” *Jurnal Algoritma*, vol. 21, no. 1, pp. 119–130, May 2024.
- [11] W. Handayani, “Analisis Pengelompokan Tingkat Kepuasan Masyarakat Terhadap Kinerja Korem 031/Wir Bima Menggunakan K-Means Clustering Berbasis Web - Repository Universitas Lancang Kuning,” *Unilak.ac.id*, 2023.
- [12] Zein, A., Farizy, S., & Suharyanto, E. (2022). "Sentimen Analisis Pada Komentar Pendek Evaluasi Dosen Oleh Mahasiswa (Edom) Program Studi Sistem Informasi Universitas Pamulang," *Jurnal Ilmu Komputer*, 5(01), 17–23.