**Cyberspace:** Jurnal Pendidikan Teknologi informasi
Volume 10, Nomor 2, Oktober 2026, hal. 33 – 58
ISSN 2598-2079 (print) | ISSN 2597-9671 (online)

# MULTIMODAL FUSION OF LIKERT FEATURES AND TEXT SENTIMENT FOR EARLY LECTURER PERFORMANCE PREDICTION

**Basrul**[1], **Khairan Ar**[2]

[1]Universitas Islam Negeri Sultanah Nahrasiyah Lhokseumawe
[2]Universitas Islam Negeri Ar-Raniry Banda Aceh
E-mail: [1]basrul.a.majid@uinsuna.ac.id

**Abstract**

Student evaluation of teaching (SET) is a central instrument for quality assurance and human-resource development in higher education, yet two methodological weaknesses persist: most analyses are cross-sectional (a single semester) and unimodal (numeric ratings analyzed in isolation from textual feedback). This study addresses both gaps using a large nine-semester corpus (2021–2025) of SET reports from an Indonesian Islamic higher education institution, comprising 11,252 valid evaluation blocks, and 44,268 free-text student comments. Because the institutional instrument changed from a four-point to a five-point Likert scale mid-period, all numeric indicators were harmonized into a common 0–100 satisfaction index. We pursue three objectives: (i) characterizing longitudinal patterns and the divergence between numeric ratings and textual sentiment; (ii) discovering performance-trajectory archetypes through feature-based K-Means clustering; and (iii) forecasting next-semester relative under-performance through a multimodal late-fusion framework benchmarked against unimodal and persistence baselines. Results reveal a pronounced ceiling effect (median index 93.6; 40% of blocks above 95) and a weak rating–sentiment correlation (Spearman 0.09), with 33.3% of top-rated blocks still containing critical comments—evidence of a halo/leniency bias that numeric scores alone conceal. Clustering identified three archetypes (Stable-High, Gradual-Decline, Volatile-At-Risk). The three-cluster solution was preferred over the statistically sharper two-cluster solution (silhouette 0.277 vs. 0.390). It only isolates the Gradual-Decline group, the most relevant target for early developmental support. For forecasting, harmonized Likert features carried the strongest signal (AUC 0.693), whereas text sentiment added limited predictive—but substantial diagnostic—value. The findings argue for treating textual feedback as a complementary diagnostic channel rather than a redundant one, and provide an operational basis for an early-warning system in academic quality assurance, positioned strictly as a formative, developmental support tool under human oversight rather than a summative or punitive evaluation instrument.

***Keywords:*** *K-Means; educational data mining; student evaluation of teaching; sentiment analysis; multimodal fusion; trajectory clustering; quality assurance*

**Abstrak**

Evaluasi Dosen oleh Mahasiswa (EDOM) merupakan instrumen utama penjaminan mutu dan pengembangan sumber daya manusia di perguruan tinggi. Namun ada dua kelemahan metodologis yang masih bertahan. Sebagian besar analisis bersifat lintas-seksi (satu semester) dan unimodal (skor numerik dianalisis terpisah dari umpan balik teks). Studi ini dapat menutup kedua celah tersebut dengan menggunakan dataset sembilan semester (2021–2025) dari sebuah Perguruan Tinggi Keagamaan Islam, yang mencakup 11,252 blok evaluasi valid, dan 44,268 komentar bebas

mahasiswa. Seluruh indikator numerik harus diharmonisasikan ke dalam indeks kepuasan 0–100 yang sebanding lantaran instrumen berubah dari skala Likert empat poin menjadi lima poin di tengah periode. Tiga target yang akan dicapai dalam riset ini adalah (i) mengarakterisasi pola longitudinal dan divergensi antara skor numerik dan sentimen teks; (ii) menemukan arketipe trajektori kinerja melalui klasterisasi K-Means berbasis fitur; serta (iii) memprediksi penurunan kinerja relatif semester berikutnya melalui kerangka fusi multimodal late-fusion yang dibandingkan terhadap basis unimodal dan persistensi. Hasil menunjukkan efek langit-langit yang tegas (median indeks 93.6; 40% blok di atas 95) dan korelasi rating–sentimen yang lemah (Spearman 0.09), dengan 33,3% blok berskor tertinggi tetap memuat komentar kritis—bukti bias halo/kelonggaran yang disembunyikan oleh skor numerik. Klasterisasi mengidentifikasi tiga arketipe (Stabil-Tinggi, Menurun Bertahap, Fluktuatif-Berisiko). Solusi tiga klaster tersebut diprioritaskan dari solusi dua klaster yang secara statistik lebih signifikan (silhouette 0,277 vs. 0,390) karena hanya solusi ini yang dapat memisahkan kelompok Menurun Bertahap sebagai sasaran pembinaan dini yang paling relevan. Fitur Likert yang telah diselaraskan dapat memberikan sinyal prediksi paling kuat (AUC 0,693), sementara sentimen teks menambah nilai prediktif yang terbatas namun bernilai diagnostik tinggi. Temuan ini menganjurkan perlakuan terhadap umpan balik teks sebagai kanal diagnostik komplementer, bukan redundan, dan menyediakan dasar operasional bagi sistem peringatan dini dalam penjaminan mutu akademik, yang diposisikan secara tegas sebagai alat pendukung pembinaan formatif di bawah pengawasan manusia, bukan instrumen evaluasi sumatif atau punitif.

***Kata kunci:*** *K-Means, penambangan data pendidikan; evaluasi dosen oleh mahasiswa; analisis sentimen; fusi multimodal; klasterisasi trajektori; penjaminan mutu*

## 1. Pendahuluan

Evaluasi Dosen oleh Mahasiswa (EDOM) atau dalam literatur internasional dikenal sebagai *Student Evaluation of Teaching* (SET) telah lama menempati posisi sentral dalam penjaminan mutu pembelajaran di perguruan tinggi. Instrumen ini memiliki fungsi yang bersifat multidimensional, tidak hanya sebagai umpan balik formatif bagi dosen, tetapi juga sebagai bukti pendukung dalam proses akreditasi program studi dan institusi serta sebagai salah satu rujukan untuk pengambilan keputusan kepegawaian dan perencanaan pengembangan kapasitas akademik [1], [2]. Di Indonesia, penyelenggaraan EDOM secara periodik melekat pada Sistem Penjaminan Mutu Internal (SPMI) dan menjadi komponen yang diawasi dalam instrumen akreditasi nasional. Akumulasi datanya dari semester ke semester membentuk aset informasi yang besar namun aset tersebut jarang diolah melampaui pelaporan rata-rata sederhana. Pemanfaatannya masih sering terbatas pada pelaporan statistik deskriptif dan agregasi nilai rata-rata, sementara komentar tertulis mahasiswa jarang dianalisis secara sistematis, sehingga potensi analitis data longitudinal tersebut belum sepenuhnya dimanfaatkan [3], [4].

Namun, keterbatasan dari implementasi tersebut setidaknya berkaitan dengan dua celah metodologis utama. Pertama, evaluasi kinerja dosen cenderung menggunakan pendekatan lintas seksi (*cross-sectional approach*) yang menempatkan hasil evaluasi pada satu titik waktu, umumnya satu semester. Pendekatan ini membatasi kemampuan sistem penjaminan mutu dalam mengenali perubahan kinerja secara temporal, termasuk pola peningkatan bertahap, penurunan berkelanjutan, maupun fluktuasi kinerja dari waktu ke waktu. Kedua, analisis EDOM masih cenderung bersifat unimodal (*unimodal analysis*), yaitu memperlakukan skor numerik berskala Likert dan komentar kualitatif

mahasiswa sebagai sumber informasi yang terpisah. Skor numerik dapat dipengaruhi oleh adanya *ceiling effect* (nilai kepentok di batas maksimal) dan *leniency/halo bias* (kecenderungan memberi nilai bagus karena merasa segan atau terlanjur berkesan positif). Jadi, nilai angka yang tinggi tidak selalu berarti mahasiswa benar-benar puas atau sama sekali tidak punya kritik [2]. Sementara itu, pengolahan data komentar teks dalam jumlah besar apabila tidak didukung oleh teknik otomatisasi tentu membutuhkan waktu dan tenaga yang signifikan [5].

Keterbatasan tersebut memiliki implikasi langsung terhadap kemampuan institusi dalam melakukan diagnosis mutu secara lebih dini. Ketika dimensi temporal tidak diperhatikan dan informasi numerik serta tekstual dianalisis secara terpisah, kritik spesifik yang tersembunyi di balik skor kuantitatif yang tinggi berpotensi tidak teridentifikasi. Kondisi tersebut dapat menyebabkan sistem penjaminan mutu lebih berorientasi pada status capaian saat ini daripada mendeteksi perubahan pola kinerja dan potensi kebutuhan pembinaan secara lebih awal. Perkembangan *Educational Data Mining* (EDM), *Natural Language Processing* (NLP), dan *Machine Learning* telah memberikan peluang untuk mengatasi keterbatasan tersebut melalui analisis data pendidikan yang lebih terstruktur dan otomatis. Penelitian terdahulu telah menunjukkan pemanfaatan analisis sentimen dan klasifikasi teks untuk mengolah umpan balik mahasiswa, mulai dari algoritma pembelajaran mesin konvensional hingga pendekatan deep learning berbasis aspek [5], [6], [7].

Dalam konteks EDOM di Indonesia, penelitian sebelumnya telah memanfaatkan algoritma *K-Means clustering* untuk memprofilkan kinerja dosen berdasarkan data evaluasi mahasiswa [8], serta menerapkan *Naïve Bayes* dan *Support Vector Machine* untuk analisis sentimen terhadap umpan balik mahasiswa [9]. Meskipun demikian, studi-studi tersebut masih menunjukkan keterbatasan yang konsisten. Sebagian penelitian menggunakan dataset yang relatif terbatas dan hanya mencakup satu hingga dua semester, sebagian lainnya masih berfokus pada satu modalitas, baik data numerik maupun teks, sedangkan kajian yang secara eksplisit menguji divergensi atau inkonsistensi antara penilaian numerik dan sentimen teks mahasiswa dalam perspektif longitudinal masih sangat terbatas [10]. Celah tersebut menunjukkan perlunya suatu kerangka analitis yang tidak hanya mempertahankan dimensi temporal, tetapi juga mengintegrasikan informasi numerik dan tekstual dalam satu proses analisis. Penelitian ini merespons kebutuhan tersebut dengan memanfaatkan data EDOM berskala besar yang mencakup sembilan semester berturut-turut (2021–2025) pada salah satu institusi PTKIN di Indonesia. Tantangan penting dalam analisis longitudinal ini adalah adanya perubahan instrumen internal institusi dari skala Likert 4-poin menjadi 5-poin di tengah periode pengamatan. Untuk mempertahankan data historis dan memungkinkan perbandingan antarsemester secara konsisten, penelitian ini menerapkan prosedur normalisasi indeks terharmonisasi pada skala 0–100, sehingga perbedaan panjang skala dapat dikendalikan tanpa mengeliminasi observasi historis. Prinsip ini sejalan dengan temuan [11] bahwa skor dari skala rating dengan panjang berbeda dapat dibuat setara melalui transformasi linear sederhana.

Berdasarkan konteks dan kesenjangan penelitian tersebut, penelitian ini diarahkan untuk menjawab tiga pertanyaan penelitian. RQ1: Bagaimana karakteristik longitudinal skor EDOM terharmonisasi, dan seberapa besar divergensi antara penilaian numerik Likert dengan sentimen teks umpan balik mahasiswa sebagai indikasi adanya *halo effect*? RQ2: Pola atau arketipe tren lintasan kinerja (*trajectory archetypes*) apa saja yang terbentuk secara otomatis dari rekam jejak longitudinal dosen menggunakan algoritma

*feature-based K-Means clustering*? RQ3: Sejauh mana kerangka kerja fusi multimodal (*multimodal late-fusion*) mampu memprediksi risiko penurunan kinerja relatif (*relative under-performance*) dosen pada semester berikutnya dibandingkan dengan baseline unimodal dan persistensi?

Untuk menjawab pertanyaan tersebut, penelitian ini memberikan empat kontribusi utama. Pertama, kontribusi empiris berupa penyediaan analisis EDOM longitudinal berskala besar selama sembilan semester pada lingkungan PTKIN. Kedua, kontribusi metodologis melalui penerapan prosedur harmonisasi indeks untuk mengintegrasikan data evaluasi yang menggunakan kombinasi skala Likert yang berbeda dalam suatu rangkaian longitudinal. Ketiga, kontribusi diagnostik melalui kuantifikasi divergensi antara rating dan sentimen sebagai parameter operasional untuk mengidentifikasi indikasi *leniency bias* dalam evaluasi mahasiswa. Keempat, kontribusi aplikatif melalui perumusan kerangka *multimodal late-fusion* dengan protokol pemisahan temporal yang ketat sebagai landasan teknis bagi pengembangan *early-warning system* pada unit penjaminan mutu perguruan tinggi yang diposisikan sebagai alat pendukung pembinaan formatif, bukan instrumen evaluasi sumatif. Konteks PTKIN juga turut memberikan dimensi kultural yang khas bagi penelitian ini. Hal ini dikarenakan norma kesantunan dan penghormatan hierarkis terhadap dosen dalam budaya akademik keagamaan dianggap dapat memperkuat kecenderungan mahasiswa dalam memberikan penilaian numerik yang longgar. Meskipun kritik tetap disampaikan secara implisit melalui umpan balik teks, sehingga menjadikan PTKIN sebagai konteks yang relevan untuk menguji divergensi rating–sentimen yang menjadi fokus penelitian ini.

## 2. Tinjauan Pustaka

### Evaluasi Kinerja Dosen dan Bias Penilaian

Evaluasi Dosen oleh Mahasiswa (EDOM) atau *Student Evaluation of Teaching* (SET) merupakan instrumen utama yang digunakan perguruan tinggi untuk mengukur efektivitas pembelajaran serta kualitas pengajaran akademisi. Konstruk EDOM secara tradisional dirancang untuk memetakan beberapa dimensi pengajaran, seperti penguasaan materi, metode penyampaian, keadilan penilaian, dan interaksi di dalam kelas. Dalam kajian klasiknya, Marsh mengargumenkan bahwa evaluasi mahasiswa memiliki reliabilitas psikometrik yang memadai, stabilitas temporal yang tinggi, serta mencerminkan konstruk multidimensional yang relatif tahan terhadap variabel pengganggu. Meskipun demikian, literatur terkini secara konsisten mengidentifikasi keterbatasan fundamental pada nilai rerata numerik EDOM akibat keberadaan bias sistematis [1], [12]. Terdapat tiga fenomena psikometrik utama yang sering merusak validitas EDOM, yaitu 1) Efek halo (*halo effect*) yang merupakan kecenderungan mahasiswa memberikan nilai seragam pada seluruh butir kuesioner berdasarkan kesan umum (*global impression*) terhadap dosen. 2) Bias kelonggaran (*leniency bias*) yang merupakana kaitan erat antara kemudahan standar penilaian nilai ujian dan skor EDOM yang tinggi. 3) Kompresi distribusi yaitu penumpukan nilai pada batas atas skala Likert yang memicu penyempitan varians data (*ceiling effect*).

Penyempitan varians dan pemusatan distribusi tersebut menyebabkan skor agregat Likert kurang sensitif dalam membedakan variasi mutu pengajaran yang mendasar. Riset [13] menunjukkan melalui simulasi bahwa sekalipun instrumen EDOM memenuhi asumsi reliabilitas tinggi, penggunaan rerata numerik untuk membandingkan atau memeringkatkan dosen secara berpasangan tetap memiliki risiko kesalahan keputusan (*false positive/negative*) yang signifikan. Oleh karena itu, para peneliti

merekomendasikan transisi dari ketergantungan tunggal pada angka kuantitatif menuju pendekatan komplementer yang turut menganalisis komentar tertulis mahasiswa secara sistematis [14]. Penggabungan data kuantitatif dengan umpan balik tekstual melalui teknik penambangan teks dapat memberikan gambaran kinerja pedagogis yang lebih holistik, termasuk menghasilkan masukan perbaikan yang lebih spesifik dan dapat ditindaklanjuti [15].

## Analisis Sentimen Umpan Balik Mahasiswa

Analisis sentimen pada umpan balik pendidikan berperan penting dalam mengekstrak polaritas emosi dan persepsi laten yang tidak terjangkau oleh skala kuantitatif. Dalam konteks pemrosesan bahasa alami (Natural Language Processing/NLP), teknologi ini mempermudah pengolahan data tekstual dalam domain institusi pendidikan [16]. Pemrosesan teks untuk analisis sentimen, termasuk pada data berbahasa Indonesia, umumnya menempuh dua paradigma utama, yaitu pendekatan berbasis leksikon (*lexicon-based*) dan pendekatan berbasis pembelajaran mesin terawasi [17]. Pendekatan berbasis leksikon memanfaatkan kamus kata untuk menentukan nilai polaritas teks secara praktis tanpa memerlukan proses pelatihan (training), sedangkan pendekatan supervised machine learning seperti Support Vector Machine menggunakan data berlabel untuk melatih model klasifikasi agar menghasilkan tingkat akurasi dan pemahaman konteks yang lebih optimal. Pendekatan berbasis leksikon, seperti pengaplikasian kamus sentimen InSet (*Indonesian Sentiment Lexicon*), bekerja dengan menghitung bobot nilai sentimen dari kata-kata yang terkandung dalam dokumen teks [18]. Kelebihan dari pendekatan ini terletak pada transparansi perhitungan, efisiensi komputasi, serta tidak diperlukannya data pelatihan berlabel (*annotated dataset*). Namun, leksikon mengalami keterbatasan serius saat berhadapan dengan bahasa tidak formal, penggunaan singkatan khas mahasiswa, idiom lokal, serta kritik tersirat (*implicit feedback*) yang sering ditemui dalam evaluasi pengajaran di Indonesia.

Di sisi lain, pendekatan pembelajaran mesin terawasi memanfaatkan ekstraksi fitur seperti *Term Frequency–Inverse Document Frequency* (TF-IDF) yang memetakan bobot pentingnya suatu kata terhadap dataset [19]. Pemetaan ini dipadukan dengan pengklasifikasi statistik atau model bahasa terpra-latih (*pre-trained language models*) seperti IndoBERT [20]. IndoBERT menggunakan arsitektur *Transformer* bi-direksional yang mampu menangkap konteks semantik secara mendalam. Walaupun arsitektur berbasis *Transformer* memberikan akurasi klasifikasi sentimen yang superior, pendekatannya membutuhkan daya komputasi yang besar serta ketersediaan data pelatihan terlabel dalam jumlah memadai. Pemilihan metode idealnya menyeimbangkan antara tingkat granularitas analisis dan ketersediaan sumber daya data.

## Educational Data Mining, Klasterisasi, dan Peramalan

*Educational Data Mining* (EDM) merupakan disiplin ilmu yang memanfaatkan teknik penambangan data untuk mengeksplorasi data unik yang dihasilkan dari lingkungan pendidikan. Salah satu penerapan utamanya adalah penggunaan algoritma klasterisasi tak-terbimbing (*unsupervised clustering*) untuk mengidentifikasi pola laten atau profil kinerja dosen tanpa memerlukan label awal. Algoritme K-Means menjadi salah satu teknik paling fundamental dalam mengelompokkan data berdasarkan jarak Euclidean antar-vektor fitur [21]. Dalam analisis data longitudinal, klasterisasi yang diterapkan pada fitur lintasan (seperti rerata nilai dasar, kemiringan tren perubahan, dan volatilitas kinerja) mampu menghasilkan klasifikasi arketipe kinerja yang intuitif bagi pengambil kebijakan.

Pada pemodelan prediktif atau peramalan kinerja, algoritma pembelajaran mesin klasik seperti Regresi Logistik [22], Random Forest [23] dan Gradient Boosting [24] tetap menjadi standar acuan yang sangat tangguh.

Penggunaan kombinasi algoritma pembelajaran mesin klasik terbukti efektif dalam memprediksi performa dan indikator akademik di tingkat perguruan tinggi [25]. Random Forest menggabungkan teknik *ensemble bootstrap aggregation* (bagging) untuk mengurangi varians, sedangkan Gradient Boosting membangun pohon keputusan secara sekuensial untuk meminimalkan fungsi kerugian (*loss function*). Pada struktur data yang bersifat teragregasi—di mana ukuran sampel terbatas pada tingkat kelas atau dosen—model ensemble klasik ini kerap memberikan kinerja yang lebih stabil dan terhindar dari *overfitting* dibandingkan model *deep learning* yang membutuhkan data masif. Hal krusial yang sering diabaikan dalam EDM prediktif adalah skema validasi data temporal. Untuk memprediksi kinerja masa depan berdasarkan data historis, penerapan *cross-validation* acak standar dapat memicu kebocoran data (*data leakage*) karena informasi dari masa depan secara tidak sengaja terikut dalam proses pelatihan model. Studi [26] menunjukkan bahwa evaluasi model pada data deret waktu perlu menjaga struktur temporal agar tidak terjadi kebocoran informasi dari periode yang lebih baru. Oleh karena itu, penelitian ini menerapkan pemisahan data temporal (*temporal splitting*), yaitu melatih model pada periode awal dan mengujinya pada periode berikutnya, sebagai skema evaluasi yang paling sesuai dengan skenario penggunaan nyata.

### Fusi Multimodal dan Posisi Penelitian

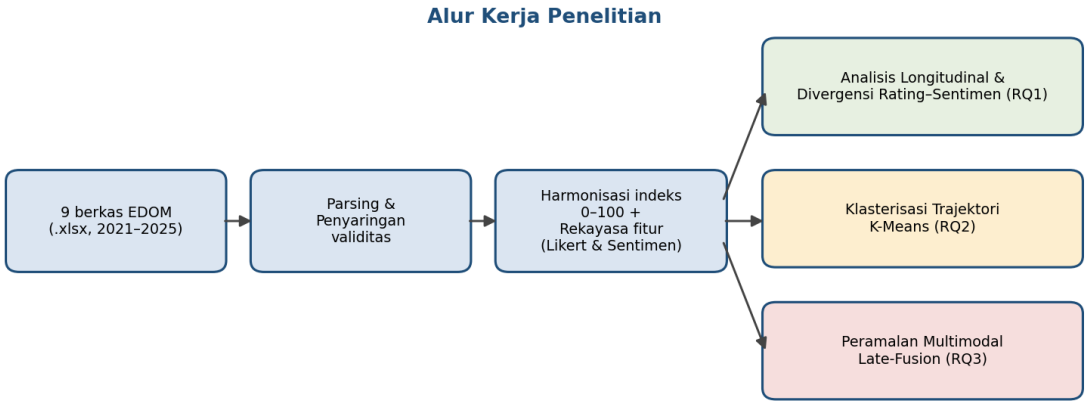
Fusi multimodal (*multimodal fusion*) adalah kerangka kerja yang memadukan informasi dari berbagai bentuk atau modalitas data heterogen untuk meningkatkan akurasi inferensi sistem. Berdasarkan taksonomi pembelajaran multimodal yang dirumuskan [27], strategi penggabungan data secara garis besar dibagi menjadi *early fusion* (fusi tingkat fitur) dan *late fusion* (fusi tingkat keputusan atau representasi). *Early fusion* menggabungkan seluruh vektor fitur dari modalitas yang berbeda ke dalam satu ruang vektor tunggal sebelum dimasukkan ke dalam model prediktor, sedangkan *late fusion* melatih model terpisah untuk setiap modalitas data dan baru menggabungkan estimasi probabilitas atau keputusan akhirnya. Dalam konteks evaluasi dosen, fusi multimodal memadukan indikator kuantitatif teragregasi (skala Likert) dengan fitur turunan tekstual (skor sentimen umpan balik).

Penelitian terkini lebih banyak yang cenderung langsung dalam menerapkan arsitektur *deep learning* kompleks seperti gabungan LSTM–CNN pada tingkat masukan mahasiswa individu. Namun, pada konteks operasional di mana data yang tersedia telah teragregasi pada tingkat mata kuliah, kelas, atau dosen, penggunaan model *deep learning* sering kali memicu fenomena *over-parameterization*. Studi ini mengambil posisi metodologis dengan menerapkan kerangka *late-fusion* berbasis pengklasifikasi klasik yang disesuaikan dengan karakteristik agregasi data. Melalui pendekatan ini, kontribusi serta nilai tambah kontributif dari modalitas teks dapat diuji secara transparan dan terisolasi. Langkah metodologis ini dipandang lebih teruji secara empiris dan dapat dipertanggungjawabkan dibandingkan memaksakan arsitektur *deep learning* kompleks tanpa dukungan granularitas data individual yang memadai.

## 3. Method

Penelitian ini menggunakan rancangan kuantitatif eksploratif-prediktif atas data sekunder. Alur kerja menyeluruh ditunjukkan pada Gambar 1, mencakup empat tahap:

akuisisi dan penguraian (parsing) data, harmonisasi indikator dan rekayasa fitur, analisis (deskriptif-longitudinal, klasterisasi, prediksi), serta interpretasi dan validasi.




Gambar 1. Alur kerja penelitian: dari penguraian sembilan berkas EDOM hingga tiga lapis analisis (deskriptif, klasterisasi trajektori, dan peramalan multimodal).

## Sumber dan Struktur Data

Data terdiri atas sembilan berkas laporan kuesioner EDOM (format .xlsx), masing-masing mewakili satu semester dari 2021 Ganjil hingga 2025 Ganjil. Setiap berkas tersusun atas blok-blok evaluasi; satu blok merepresentasikan satu kombinasi mata kuliah–kelas–dosen dan memuat: identitas (program studi, mata kuliah, SKS, kelas, dosen), jumlah responden, distribusi proporsi jawaban untuk setiap butir Likert yang dikelompokkan ke dalam tiga kategori—A (Perencanaan Perkuliahan), B (Proses Perkuliahan), dan C (Evaluasi Perkuliahan)—serta bagian Kritik & Saran berisi komentar bebas mahasiswa. Penting dicatat bahwa data bersifat agregat pada tingkat blok (proporsi jawaban), bukan respons per individu mahasiswa; hal ini menentukan pilihan metode pada tahap berikutnya. Sebuah parser khusus dikembangkan untuk mengurai struktur berulang ini secara otomatis.

## Penyaringan Validitas dan Harmonisasi Indikator

Blok dengan jumlah responden di bawah tiga atau tanpa distribusi jawaban yang sah dikeluarkan untuk menjaga keandalan estimasi. Untuk setiap kategori, skor pada skala asli dihitung sebagai rata-rata tertimbang nilai opsi terhadap proporsi jawabannya. Karena instrumen berubah dari empat menjadi lima poin, skor asli tidak langsung sebanding antarperiode. Oleh karena itu, diterapkan harmonisasi linier ke dalam indeks kepuasan 0–100:

$$Indeks = (skor_asli - 1) / (K - 1) \times 100,$$

dengan K adalah jumlah poin skala (4 atau 5). Indeks keseluruhan blok adalah rata-rata indeks ketiga kategori. Prosedur ini memetakan ujung terendah skala ke 0 dan tertinggi ke 100, sehingga perbandingan lintas-instrumen menjadi sah secara proporsional.

## Rekayasa Fitur Teks (Sentimen)

Komentar diproses melalui case folding, pembersihan simbol non-alfabet, dan tokenisasi. Karena keterbatasan ketersediaan model kontekstual luring pada lingkungan komputasi yang digunakan, skor sentimen dihitung dengan pendekatan berbasis leksikon yang dirancang khusus untuk umpan balik akademik berbahasa Indonesia: leksikon

polaritas positif dan negatif, penanganan negasi (mis. ‘tidak jelas’ membalik polaritas), pengenalan idiom non-keluhan (mis. ‘tidak ada kritik’ diperlakukan netral), serta penanda saran konstruktif (mis. ‘sebaiknya’, ‘semoga lebih’) sebagai sinyal negatif lembut yang mencerminkan budaya kritik tersirat. Setiap komentar memperoleh polaritas positif/netral/negatif, lalu diagregasi ke tingkat blok dan dosen-semester sebagai rata-rata sentimen, proporsi komentar negatif, proporsi komentar positif, dan jumlah komentar. Untuk model berbasis teks, fitur TF-IDF (unigram–bigram, maksimum 1.500 fitur) turut diekstraksi.

Pemilihan pendekatan berbasis leksikon dibandingkan model kontekstual terlatih seperti IndoBERT didasarkan pada tiga pertimbangan. Pertimbangan pertama adalah ketiadaan data anotasi sentimen berlabel manusia yang memadai untuk pelatihan atau penyesuaian (fine-tuning) terhadap 44.268 komentar yang tersedia. Kedua, kebutuhan akan transparansi perhitungan dan kemudahan implementasi tanpa ketergantungan pada infrastruktur GPU yang relevan bagi replikasi oleh unit penjaminan mutu di kampus dengan sumber daya komputasi terbatas. Ketiga, kesesuaian dengan posisi metodologis late-fusion penelitian ini, yang menggabungkan sejumlah fitur terstruktur dan berbasis teks pada tingkat dosen-semester dengan jumlah observasi yang terbatas. Penambahan representasi kontekstual berdimensi tinggi pada tahap ini berisiko menimbulkan over-parameterization relatif terhadap ukuran data yang tersedia, sejalan dengan posisi late-fusion yang telah dijelaskan pada bagian sebelumnya.

**Klasterisasi Trajektori dan Penentuan Jumlah Klaster (RQ2)**

Untuk menjawab RQ2, setiap dosen yang memiliki data pada minimal enam dari sembilan semester direpresentasikan oleh tiga fitur lintasan, yaitu level (rerata indeks kepuasan terharmonisasi), kemiringan tren (laju perubahan indeks per semester), dan volatilitas (besaran fluktuasi indeks antarsemester). Selanjutnya data dikelompokkan menggunakan algoritma K-Means. Karena tujuan klasterisasi dalam penelitian ini bersifat operasional, yaitu menghasilkan kelompok yang dapat ditindaklanjuti oleh unit penjaminan mutu, jumlah klaster tidak ditentukan semata-mata oleh koefisien silhouette tertinggi. Variasi nilai k dievaluasi secara bersamaan dengan empat kriteria: (1) kualitas pemisahan statistik yang diukur dengan koefisien silhouette, dengan syarat struktur klaster tidak terlalu lemah untuk diinterpretasikan; (2) keterbedaan substantif, yaitu setiap klaster harus berbeda secara bermakna pada sekurang-kurangnya satu dimensi lintasan (level, arah perubahan, atau stabilitas); (3) ketertindaklanjutan, yaitu setiap klaster harus dapat dipetakan ke bentuk pembinaan yang berbeda; dan (4) kecukupan ukuran klaster, agar tidak ada kelompok yang terlalu kecil untuk dijadikan dasar program pembinaan.

Pelabelan Variabel Target Risiko Penurunan Kinerja (RQ3)

Variabel target biner risiko penurunan kinerja relatif (relative under-performance) untuk dosen d pada semester t+1 dibentuk dari indeks kepuasan terharmonisasi dosen tersebut pada dua semester berturut-turut, dengan formulasi:

$$y(d, t+1) = 1, \text{ jika } Indeks(d, t+1) < Indeks(d, t); \; y(d, t+1) = 0, \text{ jika } Indeks(d, t+1) \geq Indeks(d, t),$$

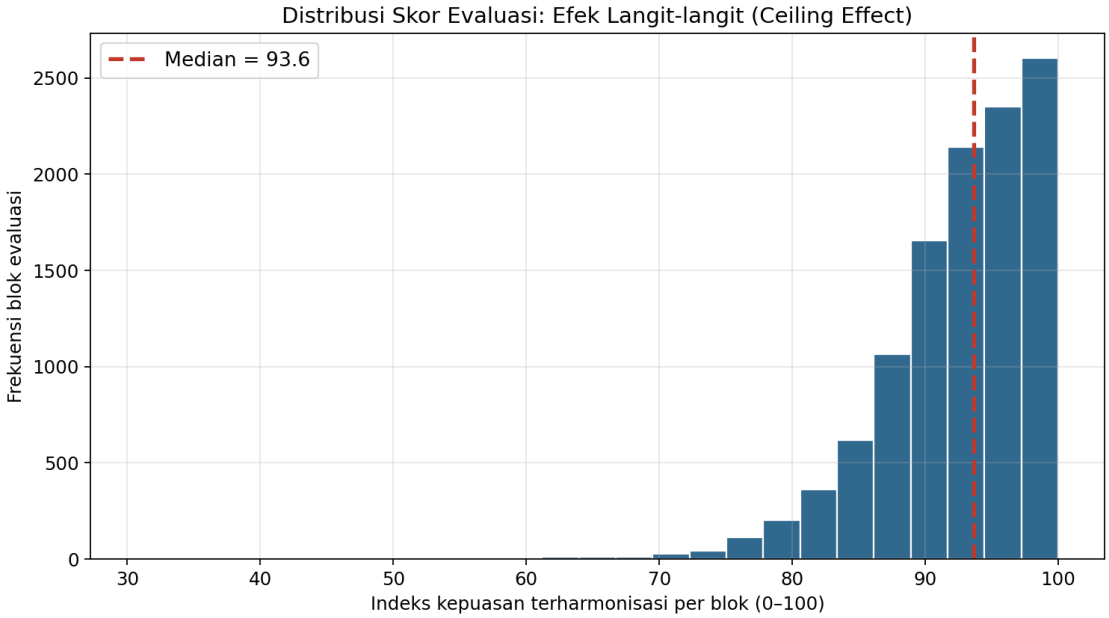
dengan Indeks (d, t) adalah indeks kepuasan terharmonisasi dosen d pada semester t. Definisi ini konsisten dengan rumusan RQ3 mengenai "risiko penurunan kinerja relatif": arah perubahan dinilai relatif terhadap performa dosen itu sendiri pada semester

sebelumnya, bukan terhadap ambang absolut tertentu, sehingga secara otomatis mengendalikan perbedaan level dasar antar-dosen. Sebagai perluasan pada penelitian mendatang, ambang berbasis simpangan baku individual dosen (misalnya y=1 jika penurunan melebihi 0,5 kali simpangan baku indeks dosen yang bersangkutan) dapat dieksplorasi untuk membedakan penurunan yang secara statistik bermakna dari fluktuasi acak berskala kecil, mengingat simpangan baku keseluruhan indeks yang relatif rendah (6,1 poin) akibat efek langit-langit.

## 4. Results

Efek Langit-Langit pada Distribusi Skor Evaluasi

Distribusi indeks kepuasan agregat sangat condong ke kiri dengan median 93,6 dari skala 0–100. Sebanyak 73% blok evaluasi bernilai ≥90 dan 40% di antaranya bahkan ≥95, sementara simpangan baku keseluruhan hanya 6,1 poin. Pola ini merupakan manifestasi efek langit-langit (ceiling effect): instrumen kehilangan daya diskriminatif pada ujung atas skala karena hampir seluruh dosen dinilai ‘sangat baik’ (Gambar 2).




Gambar 2. Distribusi indeks kepuasan terharmonisasi per blok evaluasi, memperlihatkan efek langit-langit yang tegas.

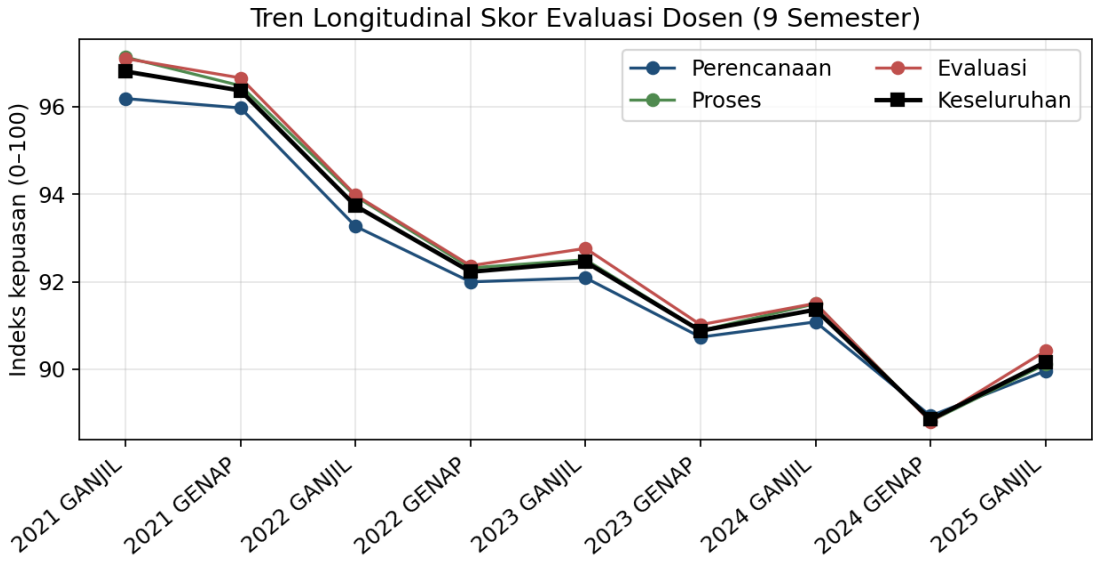
Kompresi distribusi ini dapat dijelaskan melalui tiga mekanisme yang saling menguatkan. Pertama, secara psikometrik, keterbatasan rentang (range restriction) menyebabkan varians antar-dosen yang tersedia untuk membedakan mutu menjadi kecil, sehingga reliabilitas diferensial menurun. Kedua, secara psikologi sosial, efek halo bekerja sebagai heuristik penilaian global yang diturunkan ke seluruh butir tanpa diferensiasi, sehingga skor antarbutir terkompresi dan berkorelasi tinggi. Ketiga, secara kontekstual-kultural, norma penghormatan terhadap pengajar pada perguruan tinggi keagamaan Islam (PTKI) diduga memperkuat kecenderungan pemberian skor tinggi, terutama ketika anonimitas responden tidak sepenuhnya meyakinkan.

Temuan ini memperkuat sekaligus memperluas simulasi komputasional [13] yang menunjukkan bahwa selisih skor SET yang kecil tidak dapat diandalkan untuk

mengidentifikasi pengajar yang lebih baik dalam perbandingan berpasangan. Keterbatasan rentang yang teramati pada penelitian ini bahkan lebih ketat daripada kondisi ideal yang diasumsikan dalam simulasi tersebut, sehingga menyediakan bukti empiris berskala besar bahwa kritik atas validitas SET yang disintesis [2] dan ditegaskan kembali [28] berlaku juga pada konteks PTKI, konteks yang sebelumnya jarang diteliti dalam literatur validitas SET. Perlu digarisbawahi bahwa hipotesis bias kelonggaran penilaian sebagaimana diuji [29] tidak dapat diperiksa dalam penelitian ini karena data nilai akhir mahasiswa tidak tersedia; simpulan karenanya dibatasi pada gejala kompresi distribusi, bukan pada penyebab spesifiknya, dan membuka agenda validasi lanjutan.

Tren Longitudinal Sembilan Semester

Indeks kepuasan keseluruhan menurun secara bertahap dari 96,80 pada semester 2021 Ganjil menjadi 90,16 pada 2025 Ganjil, penurunan kumulatif 6,6 poin indeks (Gambar 3). Ketiga kategori penilaian (Perencanaan Perkuliahan, Proses Pembelajaran, dan Evaluasi) bergerak hampir paralel, dengan kategori Perencanaan Perkuliahan berada paling rendah pada 8 dari 9 semester yang diamati.




*Gambar 3. Tren longitudinal indeks kepuasan keseluruhan dan per kategori selama sembilan semester*

Penurunan indeks selama periode pengamatan tidak berlangsung secara monoton, tetapi membentuk pola gigi gergaji. Pada tiga kesempatan, yaitu 2023 Ganjil, 2024 Ganjil, dan 2025 Ganjil, indeks meningkat dibandingkan semester genap sebelumnya, masing-masing dari 92,23 menjadi 92,45; dari 90,88 menjadi 91,36; dan dari 88,85 menjadi 90,16. Sebaliknya, seluruh transisi dari semester ganjil ke semester genap berikutnya menunjukkan penurunan. Penurunan tunggal paling tajam terjadi antara 2021 Genap dan 2022 Ganjil, yaitu sebesar 2,6 poin, dari 96,37 menjadi 93,73. Sementara itu, indeks terendah selama periode pengamatan tercatat pada 2024 Genap, yaitu 88,85. Pola tersebut menunjukkan bahwa perubahan indeks tidak hanya mencerminkan kecenderungan penurunan umum, tetapi juga mengandung variasi antarsemester yang perlu dibaca secara hati-hati.

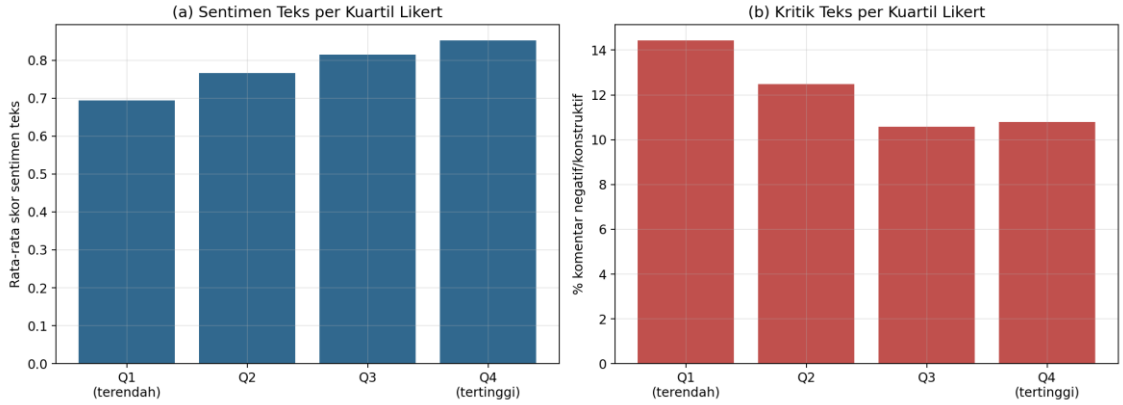
Pergerakan yang relatif paralel pada ketiga kategori menunjukkan bahwa variasi antarwaktu kemungkinan lebih banyak dipengaruhi oleh faktor tingkat institusi yang berlangsung secara serentak daripada oleh perubahan perilaku dosen secara individual.

Beberapa faktor yang mungkin berperan antara lain normalisasi ekspektasi mahasiswa pascapandemi, perluasan basis responden, dan semakin matangnya literasi evaluatif mahasiswa. Pola gigi gergaji juga mengindikasikan adanya komponen yang berkaitan dengan jenis semester, yaitu ganjil dan genap. Namun, dugaan mengenai perbedaan komposisi dan beban mata kuliah belum dapat diuji karena penelitian ini tidak memiliki data pada tingkat mata kuliah. Penelitian lanjutan dapat dilakukan oleh unit penjaminan mutu PTKIN dapat mulai melakukan pelacakan atribut mata kuliah (misalnya proporsi mata kuliah teori versus praktikum per program studi per semester). Sehingga hipotesis mengenai kontribusi komposisi beban mengajar terhadap pola gigi gergaji ini dapat diuji secara empiris pada siklus evaluasi berikutnya. Selain itu, terdapat dua faktor pengganggu (*confounding*) yang perlu diperhatikan. Pertama, jumlah program studi yang dievaluasi meningkat dari 18 menjadi 21 mulai 2024 Ganjil. Kedua, titik terendah pada 2024 Genap bertepatan dengan transisi instrumen ke skala lima poin. Kedua kondisi tersebut sejalan dengan literatur *concept drift* [30], [31] yang menunjukkan bahwa perubahan karakteristik data antarperiode dapat menurunkan keandalan model prediktif maupun interpretasi tren.

Dengan mempertimbangkan adanya faktor-faktor tersebut, temuan pada subbagian ini sebaiknya dipahami pada tingkat deskriptif dan tidak digunakan untuk menarik kesimpulan kausal mengenai perubahan kinerja individual dosen. Secara khusus, Kategori Perencanaan Perkuliahan yang secara konsisten menunjukkan nilai terendah lebih tepat dipahami sebagai indikasi persoalan struktural yang berkaitan dengan tata kelola kurikulum dan dukungan administratif institusi, bukan semata-mata sebagai cerminan kinerja individual. Oleh karena itu, implikasi perbaikannya lebih tepat diarahkan pada kebijakan di tingkat program studi atau institusi, misalnya melalui penguatan tata kelola kurikulum dan dukungan administratif, daripada semata-mata melalui pembinaan terhadap dosen secara perorangan.

Divergensi Rating–Sentimen sebagai Bukti Efek Halo

Pada 10.092 blok yang memuat sekurang-kurangnya satu komentar bersih, korelasi antara indeks kepuasan blok dan rata-rata sentimen komentarnya sangat lemah yaitu 0,126 (Pearson) dan 0,092 (Spearman). Sementara korelasi dengan proporsi komentar negatif juga lemah (Pearson −0,115; Spearman −0,113).

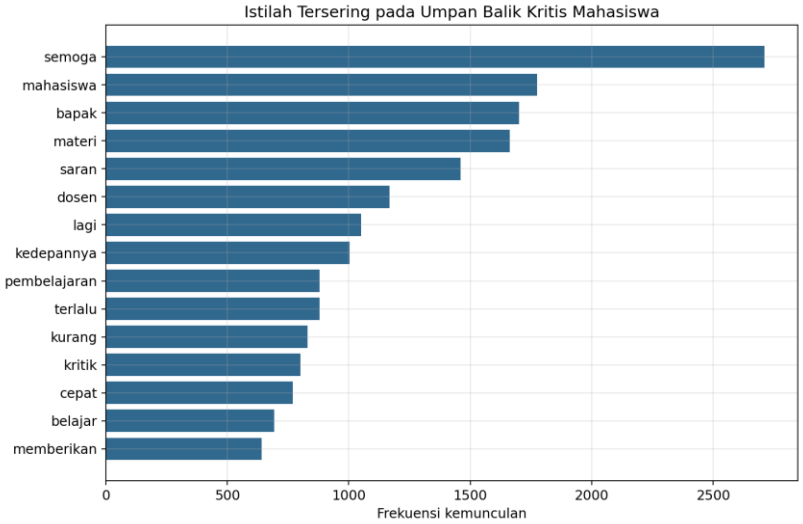



*Gambar 4. (a) Rata-rata sentimen teks per kuartil indeks Likert; (b) persentase komentar kritis pada blok berskor tinggi — bukti divergensi rating–sentimen*

Seluruh korelasi signifikan secara statistik ($p<0,001$) semata karena besarnya ukuran sampel, sehingga yang bermakna substantif adalah besaran koefisiennya yang kecil,

bukan signifikansinya. Pola pada tingkat kuartil memperjelas struktur hubungan tersebut (Gambar 4). Rata-rata skor sentimen teks meningkat monoton dari kuartil Likert terendah ke tertinggi (Q1=0,694; Q2=0,766; Q3=0,814; Q4=0,852), tetapi seluruh rentang variasi hanya 0,16 poin. Persentase komentar bernada kritis menurun dari Q1 ke Q3 (14,4%→12,5%→10,6%) namun mendatar pada Q4 (10,8%). Pada tingkat blok, penurunannya lebih tegas: proporsi blok yang memuat sekurang-kurangnya satu komentar kritis menurun dari 44,3% (Q1) menjadi 31,0% (Q4).

Analisis pada tingkat blok memperkuat simpulan ini: dari 3.798 blok berindeks sangat tinggi (≥95) yang memuat komentar bersih, sebanyak 1.266 blok (33,3%) tetap memuat sekurang-kurangnya satu komentar kritis atau konstruktif. Dengan demikian, arah hubungan antara skor Likert dan sentimen sesuai harapan teoretis, tetapi kekuatannya lemah, bukan ketiadaan hubungan, melainkan lemahnya besaran hubungan. Sehingga skor numerik tidak dapat digunakan untuk menyimpulkan ketiadaan keluhan. Analisis frekuensi istilah pada komentar kritis (Gambar 5) memberikan tekstur kualitatif atas pola tersebut. Istilah yang paling sering muncul meliputi semoga (2.712), mahasiswa (1.776), bapak (1.702), materi (1.663), saran (1.461), dan dosen (1.168). Dominasi kata semoga, saran, dan kedepannya menunjukkan bahwa kritik mahasiswa umumnya disampaikan dalam bentuk harapan yang santun, bukan celaan langsung. Pengelompokan tematik mengidentifikasi tiga keluhan berulang: kecepatan dan kejelasan penyampaian materi, beban serta arahan tugas, dan kedisiplinan waktu/kehadiran.



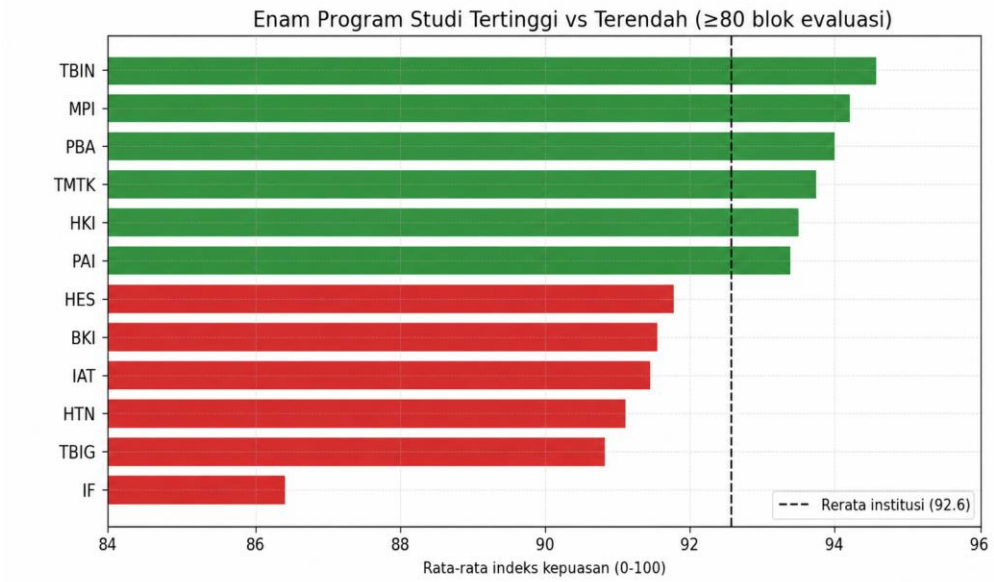

*Gambar 5. Istilah paling sering muncul pada umpan balik kritis mahasiswa*

Interpretasi yang diajukan adalah bahwa skor Likert dan komentar bebas mengukur konstruk yang berbeda, bukan konstruk yang sama dengan derajat derau berbeda: skor tinggi menandakan tiadanya ketidakpuasan global, sedangkan komentar kritis menandakan adanya hambatan spesifik, dan keduanya dapat hadir bersamaan tanpa kontradiksi logis. Temuan ini memperluas kesimpulan [4] bahwa komentar tertulis mahasiswa memuat informasi yang tidak tertangkap oleh skor rating, dengan menyediakan kuantifikasi eksplisit atas besaran ketidaksesuaian tersebut pada data longitudinal berskala besar. Temuan ini juga konsisten dengan struktur keluhan yang diidentifikasi melalui *aspect-based opinion mining* pada umpan balik mahasiswa terhadap kinerja dosen [7], yang mengindikasikan bahwa struktur keluhan mahasiswa relatif serupa lintas konteks. Dukungan konseptual dan metodologis untuk analisis

sentimen berbasis aspek serta identifikasi topik pada umpan balik pembelajaran dapat ditemukan dalam karya-karya terkait: kerangka baku tugas ABSA beserta komponen-komponennya [32], penggunaan STM untuk menambang ulasan kursus berskala besar [33], serta landasan metodologis STM untuk mengidentifikasi topik dan variasi tema berdasarkan kovariat [34]. Karena rancangan penelitian bersifat observasional, divergensi yang teramati merupakan indikasi atas bekerjanya efek halo; penegasan kausal menuntut desain eksperimental atau data tingkat-mahasiswa. Dengan demikian, seruan triangulasi metode kuantitatif dan kualitatif yang diajukan [2] dan [28] yang selama ini bersifat normatif yang memperoleh instrumen operasional dan divergensi rating–sentimen dapat dihitung secara rutin dan dilaporkan sebagai indikator mutu tersendiri.

Variasi Antar program Studi

Pada tingkat program studi (prodi), dengan ambang minimal 80 blok evaluasi, program studi tertinggi (TBIN, indeks rata-rata 94,7) dan terendah (IF, 86,4) menunjukkan rentang 8,3 poin yang setara sekitar 1,4 simpangan baku distribusi blok, sehingga secara statistik tidak dapat diabaikan meskipun seluruh prodi tetap berada pada zona kepuasan tinggi (Gambar 6). Dari 22 prodi hasil penyelarasan label, 18 memenuhi ambang minimal blok; seluruh prodi pada kelompok tertinggi berada di atas rerata institusi (92,6), dan seluruh prodi pada kelompok terendah berada di bawahnya.




*Gambar 6. Enam program studi dengan indeks kepuasan tertinggi (hijau) dan enam terendah (merah), dibatasi pada program studi dengan ≥80 blok evaluasi. Garis putus-putus menunjukkan rerata institusi (92,6).*

Variasi ini diduga bersumber dari perbedaan karakteristik disiplin ilmu dan mata kuliah, seperti tingkat abstraksi materi, rasio dosen–mahasiswa, dan intensitas praktikum. Karakteristik mata kuliah dilaporkan turut memengaruhi skor SET [2], di samping berbagai bias sosial yang disintesis oleh [35]. Interpretasi peringkat antar prodi juga perlu dilakukan secara hati-hati karena jumlah blok per prodi tidak seimbang; dua prodi yang baru dibuka selama periode pengamatan (PPG, n=20 blok; IPII, n=11 blok) tidak memenuhi ambang minimal dan karenanya tidak ditampilkan. Implikasinya, pemeringkatan antar prodi berdasarkan selisih skor semacam ini tidak layak dijadikan

dasar penilaian kinerja individual, karena rentangnya berimpitan dengan derau pengukuran sebagaimana diperlihatkan oleh efek langit-langit pada bagian sebelumnya.

Arketipe Trajektori Kinerja Dosen (RQ2)

Klasterisasi K-Means atas fitur lintasan (level rerata, kemiringan tren, dan volatilitas) diterapkan pada 235 dosen yang hadir pada sekurang-kurangnya enam dari sembilan semester. Koefisien silhouette tertinggi diperoleh pada solusi dua klaster (0,390); namun solusi tiga klaster (silhouette 0,277) dipilih karena granularitas interpretatif yang lebih relevan bagi pemangku kepentingan penjaminan mutu. Pilihan ini merupakan pertukaran eksplisit antara kualitas pemisahan statistik dan kegunaan praktis; nilai silhouette 0,277 tergolong lemah-hingga-sedang, sehingga keanggotaan klaster sebaiknya dibaca sebagai indikasi triase.

Secara lebih spesifik, solusi dua klaster hanya memisahkan kelompok ekstrem Stabil-Tinggi versus Berisiko-Rendah, sehingga gagal dalam mengidentifikasi kelompok Menurun Bertahap. Karena penurunannya belum tampak apabila hanya dibaca dari skor pada satu titik waktu. Nilai silhouette pada rentang lemah-hingga-sedang (0,2-0,4) bukan hal yang tidak lazim pada klasterisasi data perilaku atau persepsi manusia yang memiliki variansi tinggi. Dengan demikian, pemilihan k=3 mencerminkan pertukaran yang disengaja antara kemurnian statistik dan kegunaan operasional bagi pemangku kepentingan penjaminan mutu.

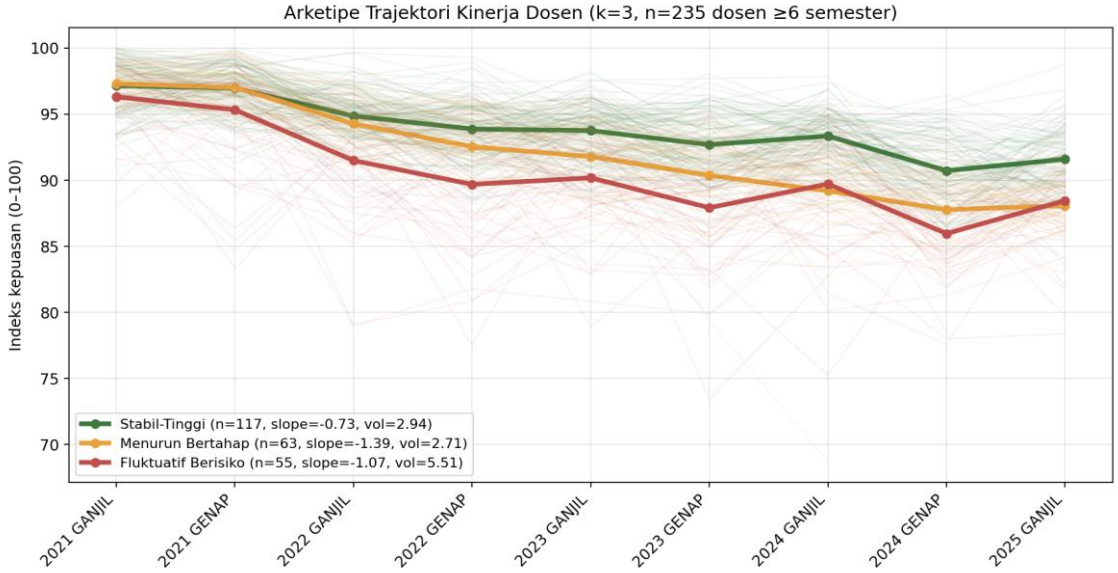
Tabel 1. Perbandingan solusi klasterisasi K-Means k=2 dan k=3 (n=235 dosen)

| Kriteria | k = 2 | k = 3 (dipilih) |
|---|---|---|
| Koefisien silhouette | 0,390 (pemisahan statistik lebih tegas) | 0,277 (lemah-hingga-sedang) |
| Arketipe yang terbentuk | Dua kutub: Stabil-Tinggi dan Berisiko-Rendah | Stabil-Tinggi (n=117), Menurun Bertahap (n=63), Fluktuatif-Berisiko (n=55) |
| Keterbedaan substantif | Terutama membedakan level kinerja (tinggi vs. rendah) | Membedakan level, arah perubahan (slope), dan stabilitas (volatilitas) |
| Deteksi penurunan bertahap pada level tinggi | Tidak terpisah sebagai kelompok tersendiri | Teridentifikasi sebagai klaster tersendiri (slope −1,39; volatilitas 2,71) |
| Ketertindaklanjutan bagi LPM | Dua kategori umum, kurang spesifik untuk pembinaan | Tiga jalur pembinaan yang berbeda (pemantauan tren, penelusuran fluktuasi, sumber praktik baik) |

Tabel 1 merangkum perbandingan kedua solusi tersebut berdasarkan kriteria yang ditetapkan pada bagian Metode. Penurunan silhouette dari 0,390 menjadi 0,277 dapat dipahami secara struktural bahwa kelompok Menurun Bertahap menempati wilayah transisi dalam ruang fitur, karena level awalnya masih tinggi seperti kelompok Stabil-Tinggi tetapi laju penurunannya paling curam (slope −1,39 per semester) dengan volatilitas paling rendah (2,71). Pemisahan kelompok transisi semacam ini sebagai klaster tersendiri secara alami memperkecil jarak antarklaster, sehingga sebagian penurunan silhouette justru merupakan konsekuensi dari tertangkapnya pola yang paling relevan secara operasional. Dari sisi fungsional, solusi tiga klaster menghasilkan kelompok dengan ukuran yang memadai (klaster terkecil n=55 atau 23,4% dari 235 dosen) dan masing-masing dapat dipetakan ke jalur pembinaan yang berbeda, yaitu pemantauan tren

dan penyegaran strategi pembelajaran untuk Menurun Bertahap, penelusuran penyebab fluktuasi dan konteks penugasan mengajar untuk Fluktuatif-Berisiko, serta pemanfaatan sebagai sumber praktik baik untuk Stabil-Tinggi.

Dari sisi operasional, dalam kerangka peringatan dini, kegagalan untuk mengenali dosen yang sedang menurun secara bertahap (false negative) lebih merugikan daripada berkurangnya ketegasan batas antar klaster. Karena kesempatan pembinaan dini akan terlewat hingga penurunan tersebut baru tampak pada skor absolut. Atas dasar pertimbangan tersebut, k=3 diprioritaskan dibandingkan k=2, dengan catatan bahwa keanggotaan klaster tetap diperlakukan sebagai indikasi awal untuk triase pembinaan, bukan label definitif atas kinerja individu.




*Gambar 7. Tiga arketipe trajektori kinerja dosen hasil klasterisasi K-Means (garis tebal = sentroid).*

Gambar 7 menunjukkan tiga arketipe trajektori kinerja dosen berdasarkan pola perubahan indeks kepuasan selama sembilan semester, dengan total 235 dosen yang memiliki sedikitnya enam semester pengamatan. Kelompok Stabil-Tinggi (n=117) merupakan kelompok terbesar dan mempertahankan tingkat kinerja paling tinggi dengan rata-rata level sekitar 93,9, penurunan yang relatif landai (−0,73 poin per semester), serta volatilitas rendah (2,94). Kelompok Menurun Bertahap (n=63) memiliki level awal yang masih tinggi, tetapi menunjukkan arah penurunan paling konsisten dengan slope −1,39 per semester. Meskipun penurunannya lebih cepat, volatilitas kelompok ini justru paling rendah (2,71), sehingga pola yang terlihat pada grafik bukan berupa perubahan yang naik-turun secara tajam, melainkan penurunan yang berlangsung secara sistematis. Sementara itu, Fluktuatif-Berisiko (n=55) memiliki level rata-rata paling rendah (90,4), slope −1,07, dan volatilitas tertinggi (5,51), yang terlihat dari garis individu yang lebih bergejolak dibandingkan dua kelompok lainnya.

Temuan tersebut menunjukkan bahwa risiko kinerja dosen tidak hanya ditentukan oleh seberapa tinggi nilai kinerja pada suatu periode, tetapi juga oleh arah perubahan dan kestabilannya dari waktu ke waktu. Kelompok Menurun Bertahap penting diperhatikan karena kinerjanya masih berada pada level relatif tinggi sehingga berpotensi tidak teridentifikasi apabila pemantauan hanya menggunakan ambang batas nilai pada satu periode. Sebaliknya, kelompok Fluktuatif-Berisiko memperlihatkan ketidakstabilan yang

lebih besar, sehingga penanganannya tidak selalu sama dengan dosen yang mengalami penurunan secara konsisten. Dengan demikian, hasil klasterisasi memberikan dasar untuk membedakan intervensi: kelompok Menurun Bertahap lebih tepat dipantau melalui monitoring tren dan penyegaran strategi pembelajaran, sedangkan kelompok Fluktuatif-Berisiko memerlukan identifikasi faktor penyebab variasi dan evaluasi konteks penugasan mengajar secara lebih mendalam.

Pendekatan ini memperluas penggunaan K-Means pada data EDOM yang sebelumnya lebih banyak digunakan untuk mengelompokkan dosen berdasarkan tingkat kinerja pada periode tertentu, seperti pada [8] dan [36]. Perbedaannya terletak pada unit informasi yang digunakan: jika klaster berbasis level terutama menjawab pertanyaan "siapa yang memiliki kinerja rendah saat ini?", maka klaster berbasis trajektori menjawab pertanyaan yang lebih dinamis, yaitu "siapa yang menunjukkan kecenderungan menuju kondisi yang lebih rendah atau semakin tidak stabil?". Pergeseran perspektif ini penting bagi penjaminan mutu karena memungkinkan identifikasi risiko dilakukan lebih dini. Dengan melihat level, arah perubahan, dan volatilitas secara bersamaan, sistem monitoring tidak hanya mendokumentasikan kondisi kinerja dosen, tetapi juga memberikan sinyal untuk menentukan prioritas pemantauan dan bentuk intervensi yang lebih sesuai.

Kinerja Peramalan Risiko Penurunan Kinerja (RQ3)

Bagian ini menunjukkan bahwa peran utama sentimen teks dalam kerangka fusi bersifat diagnostik (menjelaskan mengapa seorang dosen berisiko) dan bukan aditif secara prediktif (AUC), sebagaimana terlihat pada perbandingan berikut.

Tabel 2. Kinerja peramalan risiko penurunan kinerja semester berikutnya (data uji temporal, n=717)

| Modalitas Pengklasifikasi | Akurasi | Presisi | Recall | F1 | AUC |
|---|---|---|---|---|---|
| *Persistensi (baseline)* | *0,654** | *0,481** | *0,481* | *0,481* | *0,611* |
| Likert-saja · Regresi Logistik | 0,509 | 0,398 | 0,925* | 0,557* | 0,693* |
| Likert-saja · SVM (RBF) | 0,519 | 0,398 | 0,866 | 0,545 | 0,642 |
| Likert-saja · Random Forest | 0,608 | 0,437 | 0,611 | 0,510 | 0,642 |
| Likert-saja · Gradient Boosting | 0,559 | 0,383 | 0,527 | 0,444 | 0,613 |
| Teks-saja · Regresi Logistik | 0,628 | 0,410 | 0,268 | 0,324 | 0,531 |
| Teks-saja · SVM (RBF) | 0,605 | 0,388 | 0,318 | 0,349 | 0,526 |
| Teks-saja · Random Forest | 0,639 | 0,384 | 0,138 | 0,203 | 0,509 |
| Teks-saja · Gradient Boosting | 0,653 | 0,375 | 0,063 | 0,108 | 0,535 |
| Fusi Ramping · Regresi Logistik | 0,538 | 0,406 | 0,828 | 0,545 | 0,682 |
| Fusi Ramping · SVM (RBF) | 0,513 | 0,392 | 0,833 | 0,533 | 0,647 |
| Fusi Ramping · Random Forest | 0,608 | 0,434 | 0,573 | 0,494 | 0,643 |
| Fusi Ramping · Gradient Boosting | 0,590 | 0,409 | 0,515 | 0,456 | 0,624 |

Model peramalan diuji pada data uji temporal (n=717) dengan data latih n=1.255, membandingkan tiga konfigurasi modalitas (Likert-saja, Teks-saja, dan Fusi Ramping) di

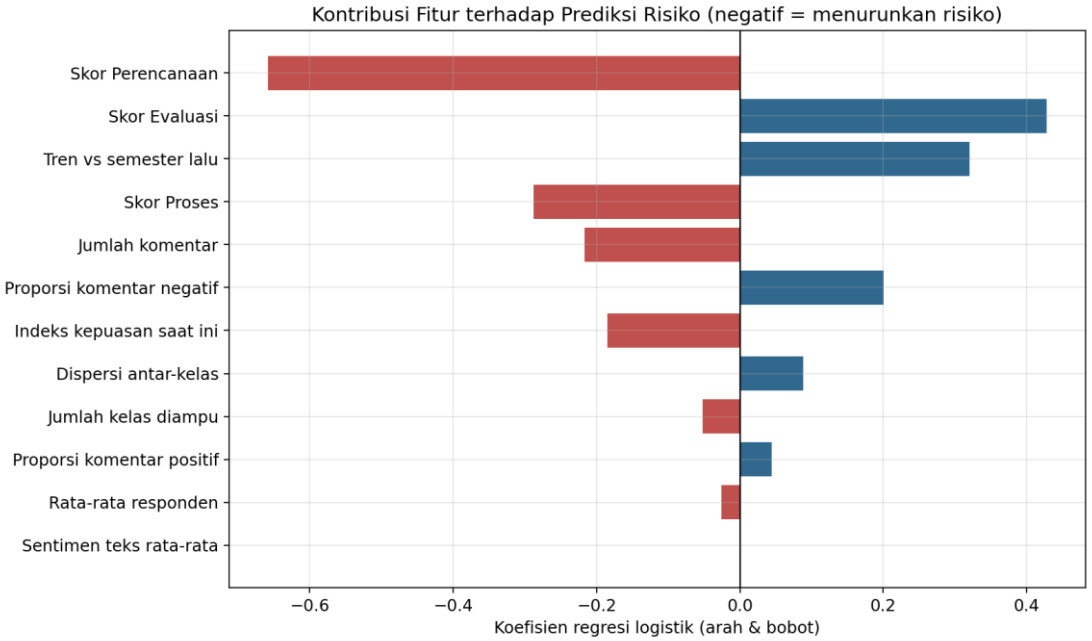
atas empat pengklasifikasi, terhadap basis Persistensi. Tabel 2 merangkum hasilnya. Empat pola utama terlihat dari Tabel 2. Pertama, model terbaik, yaitu Likert-saja dengan Regresi Logistik, menghasilkan AUC 0,693, lebih tinggi daripada baseline Persistensi sebesar 0,611, yang mengindikasikan adanya sinyal prediktif di luar ketergantungan pada kinerja sebelumnya. Kedua, fitur Likert terharmonisasi memberikan kinerja paling kuat, dengan recall 0,925, sehingga sesuai untuk fungsi penyaringan atau peringatan dini karena mampu menangkap sebagian besar dosen yang berisiko mengalami penurunan

Namun, presisi yang hanya mencapai 0,398 menunjukkan bahwa sekitar 60,2% prediksi positif merupakan false positive, sehingga model lebih tepat digunakan sebagai alat *screening* awal daripada sebagai dasar keputusan intervensi secara langsung. Ketiga, modalitas Teks-saja menunjukkan kinerja paling lemah, dengan AUC 0,509–0,535, mendekati tingkat tebakan acak; bahkan Teks-saja–Gradient Boosting hanya menghasilkan recall 0,063. Keempat, Fusi Ramping tidak memberikan peningkatan dibandingkan Likert-saja, dengan AUC tertinggi 0,682 berbanding 0,693. Lemahnya teks sebagai prediktor tunggal kemungkinan berkaitan dengan tiga kondisi: fitur teks telah diagregasi pada tingkat dosen–semester sehingga sebagian kekayaan semantiknya hilang, mayoritas komentar positif/netral membatasi variasi sinyal sentimen antar-dosen, dan terdapat perbedaan horizon temporal antara komentar yang menggambarkan kondisi saat ini dengan target yang merepresentasikan risiko pada semester berikutnya.

Tidak meningkatnya kinerja Fusi Ramping dapat dipahami sebagai indikasi adanya redundansi informasi bersyarat. Setelah riwayat skor Likert diperhitungkan, koefisien rata-rata sentimen teks pada Regresi Logistik hanya +0,001, yang menunjukkan kontribusi tambahan yang sangat kecil dari polaritas teks. Temuan ini berbeda dari [10], [37] yang menunjukkan bahwa fusi multimodal dapat meningkatkan kinerja prediksi dalam konteks pembelajaran. Perbedaan tersebut tidak harus dipandang sebagai kontradiksi, tetapi dapat dibaca sebagai syarat batas penerapan fusi multimodal. Sejalan dengan taksonomi [27] dan studi [38] manfaat fusi lebih mungkin muncul ketika modalitas menyediakan informasi yang relatif komplementer, berasal dari proses penghasil data yang berbeda, dan tersedia pada granularitas yang memadai. Dalam penelitian ini, skor Likert dan komentar teks berasal dari instrumen serta responden yang sama dan kemudian diagregasi ke tingkat dosen–semester. Dengan demikian, hasil ini memberikan indikasi bahwa penambahan modalitas tidak otomatis meningkatkan kemampuan prediksi ketika kedua modalitas membawa sinyal yang relatif beririsan.

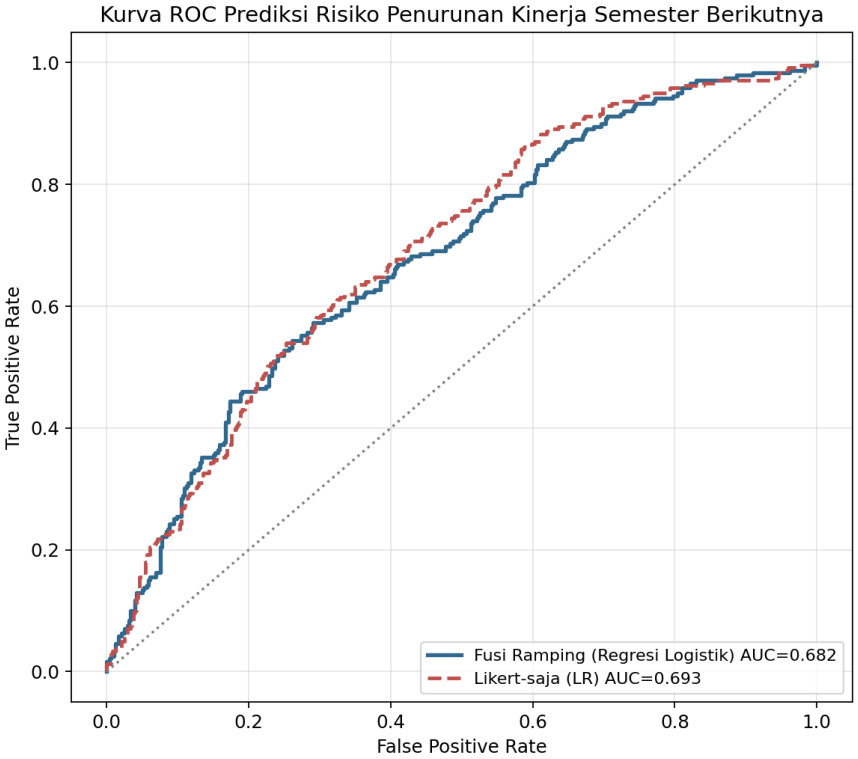
Analisis koefisien Regresi Logistik pada konfigurasi Fusi Ramping (Gambar 8) memperkuat interpretasi tersebut. Skor Perencanaan memiliki koefisien negatif terbesar (−0,658), diikuti Skor Proses (−0,288), jumlah komentar (−0,216), dan indeks kepuasan saat ini (−0,185), yang menunjukkan bahwa fitur-fitur tersebut memberikan kontribusi relatif lebih besar dalam membedakan risiko penurunan pada model. Sebaliknya, Skor Evaluasi memiliki koefisien positif terbesar (+0,428). Arah koefisien ini perlu ditafsirkan secara hati-hati karena dapat dipengaruhi oleh supresi akibat kolinearitas antarskor kategori. Koefisien tersebut sebaiknya dipahami sebagai bobot prediktif dalam konteks model, bukan sebagai hubungan kausal. Hal menariknya adalah jumlah komentar menjadi fitur berbasis teks yang relatif paling informatif, sedangkan rata-rata sentimen hanya memiliki koefisien +0,001. Temuan ini menunjukkan bahwa kontribusi teks terhadap peramalan lebih mungkin terletak pada intensitas keterlibatan responden daripada pada polaritas sentimen. Koefisien negatif terbesar pada Skor Perencanaan sejalan dengan temuan pada Bagian Tren Longitudinal.

Kategori ini secara konsisten paling rendah dan lebih mencerminkan persoalan tata kelola kurikulum serta dukungan administratif di tingkat program studi (misalnya konsistensi rencana pembelajaran semester dan penjadwalan) daripada kinerja individual dosen; dengan demikian, bobotnya yang besar dalam model peramalan turut memperkuat argumen bahwa intervensi atas dimensi ini sebaiknya diarahkan pada kebijakan tingkat program studi, bukan semata pembinaan personal. Secara keseluruhan, hasil Tabel 2 mendukung penggunaan riwayat penilaian terstruktur sebagai sinyal utama untuk peringatan dini, sementara informasi tekstual lebih tepat diposisikan sebagai sumber informasi pelengkap ketika granularitas dan independensinya terhadap data terstruktur dapat dipertahankan.




*Gambar 8. Kontribusi fitur terhadap prediksi risiko (koefisien regresi logistik; negatif = menurunkan risiko).*

Gambar 9 memperlihatkan bahwa kurva ROC model Fusi Ramping (Regresi Logistik, AUC = 0,682) dan model Likert-saja (Regresi Logistik, AUC = 0,693) saling berimpit hampir di sepanjang rentang nilai false positive rate, yang mengindikasikan bahwa penambahan fitur teks hanya memberikan nilai tambah marginal terhadap performa peramalan—bahkan sedikit lebih rendah dibandingkan model berbasis Likert saja. Secara keseluruhan, nilai AUC pada penelitian ini berada pada kategori lemah-hingga-sedang (rentang 0,509–0,693 menurut konvensi Hosmer & Lemeshow), masih di bawah capaian yang umum dilaporkan pada sejumlah studi prediksi risiko mahasiswa berbasis pembelajaran mesin, yang kerap melampaui AUC 0,80 [39], [40], [41]. Perbedaan performa ini diduga berakar dari faktor struktural data, bukan semata keterbatasan metodologis: studi-studi terdahulu tersebut umumnya memanfaatkan data aktivitas individual berfrekuensi tinggi—seperti *log clickstream* pada *Virtual Learning Environment* (VLE) dan *Learning Management System* (LMS) yang terekam setiap hari atau bahkan setiap sesi, sedangkan penelitian ini berbasis data persepsi teragregasi dengan hanya sembilan titik waktu pengukuran. Dengan demikian, sifat dan resolusi data tampaknya menjadi determinan penting bagi batas atas kinerja model, di samping pemilihan algoritmenya.




*Gambar 9. Kurva ROC: model Fusi Ramping vs Likert-saja pada data uji temporal.*

## Discussion

Seluruh tahap analisis menunjukkan pola yang konsisten mengenai keterbatasan dan kegunaan masing-masing jenis data dalam evaluasi kinerja dosen. Skor numerik tetap informatif untuk menggambarkan posisi relatif dan memperkirakan arah perubahan kinerja, tetapi efek langit-langit menyebabkan sebagian besar skor terkonsentrasi pada nilai tinggi sehingga kemampuan membedakan kinerja antar-dosen menjadi terbatas. Sebaliknya, data teks memiliki daya prediksi yang rendah ketika digunakan secara mandiri, tetapi memberikan informasi diagnostik yang tidak tersedia dalam skor numerik, terutama mengenai aspek pembelajaran yang dikeluhkan mahasiswa dan hal-hal yang perlu diperbaiki. Dengan demikian, kontribusi teks dalam penelitian ini tidak terutama berupa peningkatan akurasi prediksi, melainkan peningkatan interpretabilitas dan ketertindaklanjutan hasil analisis. Temuan ini menunjukkan bahwa keberhasilan fusi multimodal tidak seharusnya dinilai hanya dari kenaikan AUC. Pada data EDOM yang teragregasi dan menunjukkan indikasi bias halo, kedua modalitas lebih tepat dipandang sebagai sumber informasi yang saling melengkapi secara fungsi daripada sebagai fitur yang harus menghasilkan peningkatan metrik secara aditif.

Implikasi teoretis pertama berkaitan dengan divergensi antara rating dan sentimen. Penelitian ini menunjukkan bahwa perbedaan antara penilaian numerik dan isi naratif dapat diperlakukan sebagai konstruk yang dapat diukur, bukan hanya sebagai fenomena yang dijelaskan secara kualitatif. Efek langit-langit menyebabkan variasi skor numerik menjadi relatif sempit, sementara komentar mahasiswa masih dapat memuat kritik terhadap aspek tertentu. Kondisi tersebut terlihat dari korelasi Spearman yang hanya 0,092 dan temuan bahwa 33,3% blok dengan indeks rating ≥95 masih mengandung kritik. Dengan demikian, divergensi rating–sentimen dapat dihitung dan dipantau secara rutin

sebagai indikator tambahan dalam evaluasi mutu. Pendekatan ini memperkuat gagasan tentang pentingnya triangulasi sumber informasi dalam evaluasi pengajaran [28] sekaligus memberikan bentuk operasional yang lebih konkret terhadap gagasan tersebut. Pada saat yang sama, hipotesis bahwa tingginya skor EDOM berkaitan dengan grading leniency tidak dapat diuji dalam penelitian ini karena data nilai akhir mahasiswa tidak tersedia [29].

Implikasi teoretis kedua menyangkut cara menilai manfaat fusi multimodal. Hasil penelitian memperlihatkan bahwa modalitas Likert memiliki kemampuan prediksi paling kuat, dengan AUC 0,693 dibandingkan 0,611 pada baseline persistensi, sedangkan model berbasis teks hanya menghasilkan AUC 0,509–0,535 dan model fusi mencapai AUC 0,682. Temuan tersebut mendukung pembedaan antara komplementaritas fungsional dan aditivitas metrik. Sebuah modalitas dapat memberikan manfaat substantif meskipun tidak meningkatkan kinerja prediksi ketika digabungkan dengan modalitas lain. Dalam penelitian ini, teks berfungsi menjawab pertanyaan apa yang perlu diperbaiki, sedangkan skor Likert lebih berguna untuk menjawab siapa yang perlu mendapat perhatian lebih awal. Perspektif ini juga membantu menjelaskan mengapa fusi tidak selalu menghasilkan peningkatan performa seperti yang sering diasumsikan dalam penelitian multimodal [27], [37], [38]. Sekalipun penelitian lain menunjukkan bahwa fusi dapat meningkatkan akurasi prediksi apabila disertai strategi seleksi atribut atau ensemble yang tepat [10], [42]. Dalam konteks EDOM, ketidaksesuaian granularitas data dan kemungkinan ketergantungan proses pembentukan data merupakan keterbatasan yang dapat membatasi manfaat fusi meskipun kedua modalitas secara substantif memiliki informasi yang berbeda.

Implikasi ketiga berkaitan dengan dimensi longitudinal dan penerapan hasil analisis dalam sistem peringatan dini. Klasterisasi terhadap 235 dosen menghasilkan tiga arketipe, yaitu Stabil-Tinggi, Menurun Bertahap, dan Fluktuatif-Berisiko, dengan nilai silhouette 0,277 pada k=3 — nilai yang tergolong moderat dan menjadi salah satu keterbatasan metodologis penelitian ini. Solusi ini dipilih secara sadar di atas solusi k=2 (silhouette 0,390) karena hanya solusi tiga klaster yang memisahkan kelompok Menurun Bertahap sebagai sasaran pembinaan dini (Tabel 1). Meski demikian, ketiga kelompok tersebut tidak hanya membedakan tingkat kinerja, tetapi juga membedakan pola perubahan dan stabilitas kinerja, sehingga memberikan informasi yang tidak dapat diperoleh dari pengelompokan berdasarkan satu titik waktu. Temuan ini memperluas penggunaan klasterisasi K-Means dalam analisis EDOM dari pendekatan berbasis level, sebagaimana diterapkan pada studi kasus tunggal [8], menuju pendekatan berbasis lintasan longitudinal yang mampu menangkap dinamika kinerja dari waktu ke waktu. Secara praktis, hasil tersebut dapat diterjemahkan menjadi sistem peringatan dini dua lapis bagi Lembaga Penjaminan Mutu. Lapis pertama menggunakan model Likert dengan prioritas pada recall untuk menyaring dosen yang berpotensi mengalami penurunan kinerja. Lapis kedua menggunakan analisis sentimen dan aspek keluhan untuk menentukan bentuk pembinaan yang lebih spesifik. Kelompok Fluktuatif-Berisiko dapat diarahkan pada evaluasi penugasan dan konsistensi proses pembelajaran, kelompok Menurun Bertahap pada pemantauan tren dan penyegaran metode pembelajaran, sedangkan kelompok Stabil-Tinggi dapat dimanfaatkan sebagai sumber praktik baik. Pendekatan ini menempatkan prediksi sebagai alat triase dan pembinaan, bukan sebagai keputusan otomatis terhadap individu.

Secara operasional, alur dua lapis tersebut dapat diringkas sebagai berikut: (1) pada akhir setiap semester, model berbasis fitur Likert dijalankan terhadap seluruh dosen aktif

untuk menghasilkan skor risiko penurunan kinerja pada semester berikutnya; (2) dosen dengan skor risiko di atas ambang tertentu (diprioritaskan pada recall tinggi agar kasus berisiko tidak terlewat) diteruskan ke lapis kedua; (3) pada lapis kedua, komentar teks dosen yang bersangkutan dianalisis melalui pendekatan berbasis leksikon untuk mengidentifikasi kata kunci keluhan dominan dan arketipe trajektori yang paling sesuai (Stabil-Tinggi, Menurun Bertahap, atau Fluktuatif-Berisiko); dan (4) Lembaga Penjaminan Mutu menggunakan kombinasi skor risiko dan tema keluhan tersebut untuk menetapkan bentuk pembinaan yang proporsional, dengan keputusan akhir tetap berada pada penilaian manusia (*human-in-the-loop*) dan bukan pada otomatisasi penuh.

Terkait implikasi kebijakan, temuan ini menunjukkan perlunya perubahan cara hasil EDOM digunakan dalam pengambilan keputusan mutu. Pemeringkatan dosen atau program studi berdasarkan selisih skor numerik yang kecil sebaiknya tidak dijadikan dasar keputusan individual, karena keterbatasan rentang skor dapat membuat perbedaan kecil sulit dibedakan dari variasi pengukuran; bahkan selisih 8,3 poin antar program studi perlu ditafsirkan secara hati-hati dalam konteks tersebut. Temuan mengenai skor yang tinggi tetapi tetap disertai kritik juga konsisten dengan penelitian yang menunjukkan bahwa karakteristik disiplin ilmu dan bias sosial dapat memengaruhi hasil SET, serta bahwa perbandingan skor SET antar periode rentan terdistorsi oleh faktor eksternal seperti karakteristik tahun ajaran, sehingga interpretasi tren memerlukan model pengukuran yang disesuaikan [3], [35]. Lebih jauh, mengingat presisi model yang relatif rendah dan kerentanan SET terhadap bias, hasil sistem sebaiknya digunakan sebagai pemicu percakapan pembinaan yang melibatkan penilaian manusia, bukan sebagai instrumen sanksi administratif otomatis [13], [43].

## Etika Penggunaan Sistem Peringatan Dini: Formatif, Bukan Sumatif

Sistem peringatan dini yang diusulkan dalam penelitian ini dirancang dan ditegaskan semata-mata sebagai alat pendukung pembinaan yang bersifat formatif dan developmental, bukan sebagai instrumen evaluasi sumatif maupun punitif. Dalam kerangka ini, seluruh keluaran sistem, baik skor risiko, keanggotaan arketipe trajektori, maupun tema keluhan dari komentar teks, berfungsi sebagai sinyal awal untuk membuka ruang diskusi dan pendampingan pedagogis, bukan sebagai putusan atas kinerja seseorang. Penegasan ini memiliki dasar empiris dari hasil penelitian ini sendiri. Presisi model terbaik hanya 0,398, sehingga sekitar enam dari sepuluh dosen yang ditandai sebenarnya tidak mengalami penurunan kinerja pada semester berikutnya; koefisien silhouette 0,277 menunjukkan bahwa batas antarklaster tidak tegas; dan efek langit-langit serta indikasi bias halo membatasi validitas skor numerik sebagai ukuran mutu pengajaran yang presisi [2], [13]. Pemrioritasan recall yang menghasilkan banyak false positive hanya dapat dibenarkan dalam kerangka formatif, karena konsekuensi dari penandaan yang keliru berupa tawaran pendampingan, bukan sanksi. Sebaliknya, apabila keluaran yang sama digunakan untuk keputusan sumatif, kesalahan klasifikasi tersebut berisiko menimbulkan ketidakadilan sistematis terhadap individu [43], [44].

Untuk menjaga orientasi formatif tersebut, implementasi sistem perlu dibatasi oleh lima prinsip tata kelola. Pertama, pembatasan tujuan penggunaan: keluaran sistem tidak digunakan sebagai dasar keputusan kepegawaian yang bersifat sumatif, seperti promosi, kenaikan jabatan fungsional, remunerasi, perpanjangan kontrak, pemberian sanksi, maupun pemeringkatan dosen, dan ketentuan ini sebaiknya dinyatakan secara tertulis dalam kebijakan SPMI. Kedua, keputusan tetap berada pada manusia (*human-in-the-loop*): setiap sinyal risiko wajib ditelaah oleh LPM bersama pimpinan program studi

dengan mempertimbangkan konteks yang tidak terekam dalam data, seperti beban mengajar, karakteristik mata kuliah, dan perubahan instrumen, sebelum ditindaklanjuti. Ketiga, kerahasiaan dan pembatasan akses: status risiko dan label arketipe hanya dapat diakses oleh pihak yang berwenang memberikan pembinaan, tidak dipublikasikan, dan dikomunikasikan kepada dosen dengan istilah yang bernuansa pengembangan, misalnya "prioritas pendampingan" alih-alih "berisiko", untuk menghindari stigmatisasi; anonimitas mahasiswa penulis komentar juga tetap dijaga. Keempat, transparansi dan hak klarifikasi: dosen diinformasikan mengenai keberadaan sistem, data yang digunakan, serta keterbatasannya, dan diberi kesempatan untuk memberikan klarifikasi atau konteks atas sinyal yang dihasilkan. Kelima, audit keadilan secara berkala: mengingat skor EDOM dipengaruhi oleh karakteristik disiplin ilmu dan berbagai bias sosial [35], distribusi sinyal dan kinerja model perlu diaudit secara berkala lintas program studi dan kelompok dosen untuk memastikan tidak ada kelompok yang secara sistematis lebih sering ditandai.

Konsisten dengan prinsip-prinsip tersebut, tindak lanjut atas sinyal sistem diarahkan pada bentuk-bentuk dukungan, seperti mentoring sejawat dengan dosen dari kelompok Stabil-Tinggi, lokakarya pengembangan pedagogi, observasi kelas yang bersifat suportif, dan refleksi bersama atas tema keluhan mahasiswa. Sinyal yang berakar pada persoalan struktural, khususnya pada kategori Perencanaan Perkuliahan, diarahkan pada perbaikan kebijakan dan dukungan di tingkat program studi atau institusi, bukan dibebankan kepada dosen secara perorangan. Dengan demikian, keberhasilan sistem tidak diukur dari jumlah dosen yang terdeteksi, melainkan dari perbaikan mutu pembelajaran yang dihasilkan melalui pembinaan yang tepat sasaran.

## Conclusion

Penelitian ini menunjukkan bahwa skor numerik (Likert) dan data teks pada EDOM memiliki peran yang berbeda dan saling melengkapi, bukan saling menggantikan. Skor Likert terbukti memiliki daya prediksi tertinggi terhadap penurunan kinerja dosen (AUC 0,693 dibandingkan baseline persistensi 0,611), sementara data teks memberikan nilai diagnostik yang tidak dapat digantikan oleh angka semata, khususnya dalam mengidentifikasi aspek pembelajaran yang perlu diperbaiki. Analisis lintasan longitudinal terhadap 235 dosen berhasil mengidentifikasi tiga arketipe kinerja (Stabil-Tinggi, Menurun Bertahap, Fluktuatif-Berisiko), yang menjadi dasar bagi rancangan sistem peringatan dini dua lapis berbasis *human-in-the-loop* bagi Lembaga Penjaminan Mutu. Sistem tersebut ditegaskan sebagai alat pendukung pembinaan yang bersifat formatif dan developmental: keluarannya tidak dimaksudkan sebagai dasar keputusan sumatif atau punitif seperti promosi, remunerasi, maupun sanksi, melainkan sebagai pemicu dialog pembinaan dengan keputusan akhir tetap berada pada penilaian manusia.

Kontribusi utama penelitian ini bukan terletak pada klaim bahwa satu modalitas lebih unggul daripada modalitas lainnya, melainkan pada penyatuan analisis level, lintasan, dan informasi diagnostik ke dalam kerangka mutu yang lebih hati-hati. Secara spesifik, penelitian ini memperluas literatur dengan: (1) mengoperasionalkan divergensi rating–sentimen sebagai konstruk yang dapat diukur dan dipantau secara rutin; (2) membedakan secara eksplisit antara manfaat prediktif dan manfaat diagnostik dalam fusi multimodal, sehingga keberhasilan fusi tidak seharusnya dinilai hanya dari kenaikan AUC; dan (3) menunjukkan bahwa efektivitas analitik EDOM sangat dipengaruhi oleh karakteristik dan granularitas data yang digunakan.

Untuk penelitian selanjutnya disarankan beberapa arah berikut: (1) revisi instrumen EDOM untuk meningkatkan daya diskriminatif, misalnya melalui butir berjangkar perilaku (*behaviorally anchored items*), format pilihan paksa (*forced-choice*), dan pertanyaan eksplisit mengenai aspek yang paling perlu diperbaiki; (2) pengujian hipotesis grading leniency dengan melibatkan data nilai akhir mahasiswa, yang tidak tersedia dalam penelitian ini; (3) pengembangan model pengukuran yang disesuaikan untuk mengendalikan faktor eksternal antar-periode (misalnya karakteristik tahun ajaran) agar interpretasi tren kinerja lebih valid; (4) eksplorasi data dengan granularitas yang lebih halus atau strategi seleksi atribut/ensemble untuk menguji apakah manfaat prediktif fusi multimodal dapat ditingkatkan pada konteks EDOM, sebagaimana ditunjukkan berhasil pada studi lain; serta (5) uji coba implementasi sistem peringatan dini secara formatif untuk mengevaluasi dampaknya terhadap perbaikan pembelajaran serta persepsi dosen mengenai keadilan dan kebermanfaatan sistem.

## References


[1] B. Uttl, C. A. White, and D. W. Gonzalez, "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," *Stud. Educ. Eval.*, vol. 54, pp. 22–42, Sep. 2017, doi: 10.1016/j.stueduc.2016.08.007.

[2] P. Spooren, B. Brockx, and D. Mortelmans, "On the Validity of Student Evaluations of Teaching: The State of the Art," *Rev. Educ. Res.*, vol. 83, no. 4, pp. 598–642, 2013, doi: 10.3102/0034654313496870.

[3] I. Sulis, M. Porcu, and V. Capursi, "On the use of student evaluation of teaching: A longitudinal analysis combining measurement issues and implications of the exercise," *Soc. Indic. Res.*, vol. 142, no. 3, pp. 1305–1331, 2019, doi: 10.1007/s11205-018-1946-8.

[4] F. N.-A. Alhija and B. Fresko, "Student evaluation of instruction: What can be learned from students' written comments?," *Stud. Educ. Eval.*, vol. 35, no. 1, pp. 37–44, Mar. 2009, doi: 10.1016/j.stueduc.2009.01.002.

[5] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study," *Appl. Sci.*, vol. 11, no. 9, p. 3986, Apr. 2021, doi: 10.3390/app11093986.

[6] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Sentiment Analysis: Towards a Tool for Analysing Real-Time Students Feedback," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, Cyprus: IEEE, Nov. 2014, pp. 419–423. doi: 10.1109/ICTAI.2014.70.

[7] I. Sindhu, S. Muhammad Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation," *IEEE Access*, vol. 7, pp. 108729–108741, 2019, doi: 10.1109/ACCESS.2019.2928872.

[8] N. L. P. P. Dewi, I. N. Purnama, and N. W. Utami, "Penerapan data mining untuk clustering penilaian kinerja dosen menggunakan algoritma K-Means (studi kasus: STMIK Primakara)," *J. Ilm. Teknol. Inf. Asia*, vol. 16, no. 2, pp. 105–112, 2022, doi: 10.32815/jitika.v16i2.761.

[9] T. Syafrudin, A. Hermawan, and D. Avianto, "Analisis kinerja support vector machine dan naïve Bayes classifier dalam klasifikasi sentimen dan emosi pada

umpan balik mahasiswa terhadap kinerja dosen," *J. Digit*, vol. 15, no. 2, p. 34, Nov. 2025, doi: 10.51920/jd.v15i2.442.

[10] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, "A review on data fusion in multimodal learning analytics and educational data mining," *WIREs Data Min. Knowl. Discov.*, vol. 12, no. 4, p. e1458, 2022, doi: 10.1002/widm.1458.

[11] A. M. Colman, C. E. Norris, and C. C. Preston, "Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales," *Psychol. Rep.*, vol. 80, no. 2, pp. 355–362, 1997, doi: 10.2466/pr0.1997.80.2.355.

[12] H. W. Marsh, "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness," in *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, R. P. Perry and J. C. Smart, Eds., Dordrecht: Springer Netherlands, 2007, pp. 319–383. doi: 10.1007/1-4020-5742-3_9.

[13] J. Esarey and A. Valdes, "Unbiased, reliable, and valid student evaluations can still be unfair," *Assess. Eval. High. Educ.*, vol. 45, no. 8, pp. 1106–1120, 2020, doi: 10.1080/02602938.2020.1724875.

[14] E. Santhanam, B. Lynch, and J. Jones, "Making sense of student feedback using text analysis – adapting and expanding a common lexicon," *Qual. Assur. Educ.*, vol. 26, no. 1, pp. 60–69, 2018, doi: 10.1108/QAE-11-2016-0062.

[15] S. Gottipati, V. Shankararaman, and J. R. Lin, "Text analytics approach to extract course improvement suggestions from students' feedback," *Res. Pract. Technol. Enhanc. Learn.*, vol. 13, art. no. 6, 2018, doi: 10.1186/s41039-018-0073-0.

[16] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," 2023, doi: 10.48550/ARXIV.2302.04359.

[17] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[18] F. Koto and G. Y. Rahmaningtyas, "InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, 2017, pp. 391–394, doi: 10.1109/IALP.2017.8300625.

[19] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972, doi: 10.1108/eb026526.

[20] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. and 10th Int. Joint Conf. Natural Lang. Process. (AACL-IJCNLP)*, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.

[21] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.

[22] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3rd ed. John Wiley & Sons, 2013. doi: 10.1002/9781118548387.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[25] B. Yusuf, M. Qalbi, B. Basrul, I. Dwitawati, M. Malahayati, and M. Ellyadi, "Implementasi algoritma naive Bayes dan random forest dalam memprediksi prestasi akademik mahasiswa Universitas Islam Negeri Ar-Raniry Banda Aceh,"

*Cyberspace: J. Pendidik. Teknol. Inf.*, vol. 4, no. 1, p. 50, Jul. 2020, doi: 10.22373/cj.v4i1.7247.

[26] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Inf. Sci.*, vol. 191, pp. 192–213, 2012, doi: 10.1016/j.ins.2011.12.028.

[27] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[28] B. Uttl, "Student evaluation of teaching (SET): Why the emperor has no clothes and what we should do about it," *Hum. Arenas*, vol. 7, no. 2, pp. 403–437, 2024, doi: 10.1007/s42087-023-00361-7.

[29] H. W. Marsh and L. A. Roche, "Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders?," *J. Educ. Psychol.*, vol. 92, no. 1, pp. 202–228, 2000, doi: 10.1037/0022-0663.92.1.202.

[30] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, art. no. 44, 2014, doi: 10.1145/2523813.

[31] B. Sonnleitner, T. Madou, M. Deceuninck, F. Theodosiou, and Y. R. Sagaert, "Evaluation of early student performance prediction given concept drift," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100369, Jun. 2025, doi: 10.1016/j.caeai.2025.100369.

[32] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 27–35, doi: 10.3115/v1/S14-2004.

[33] L. Yang, "Mining and visualizing large-scale course reviews of LMOOCs learners through structural topic model," *PLoS ONE*, vol. 18, no. 5, p. e0284463, 2023, doi: 10.1371/journal.pone.0284463.

[34] M. E. Roberts, B. M. Stewart, and D. Tingley, "stm: An R package for structural topic models," *J. Stat. Softw.*, vol. 91, no. 2, pp. 1–40, 2019, doi: 10.18637/jss.v091.i02.

[35] T. Heffernan, "Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching," *Assess. Eval. High. Educ.*, vol. 47, no. 1, pp. 144–154, 2022, doi: 10.1080/02602938.2021.1888075.

[36] H. S. Subangkit, T. H. Taqqa, and D. I. S. Saputra, "Lecturer performance prediction based on student evaluation data using a hybrid K-Means and random forest model," *J. Penelit. Pendidik. IPA*, vol. 12, no. 1, pp. 352–358, Jan. 2026, doi: 10.29303/jppipa.v12i1.14163.

[37] H. Ouhaichi, D. Spikol, and B. Vogel, "Research trends in multimodal learning analytics: A systematic mapping study," *Comput. Educ. Artif. Intell.*, vol. 4, p. 100136, 2023, doi: 10.1016/j.caeai.2023.100136.

[38] D. Di Mitri, J. Schneider, M. Specht, and H. Drachsler, "From signals to knowledge: A conceptual model for multimodal learning analytics," *J. Comput. Assist. Learn.*, vol. 34, no. 4, pp. 338–349, 2018, doi: 10.1111/jcal.12288.

[39] M. Adnan *et al.*, "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[40] R. Z. Pek, S. T. Ozyer, T. Elhage, T. Ozyer, and R. Alhajj, "The role of machine learning in identifying students at-risk and minimizing failure," *IEEE Access*, vol. 11, pp. 1224–1243, 2023, doi: 10.1109/ACCESS.2022.3232984.

[41] H. Waheed, S.-U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: A neural network approach," *Expert Syst. Appl.*, vol. 213, p. 118868, Mar. 2023, doi: 10.1016/j.eswa.2022.118868.

[42] W. Chango, R. Cerezo, and C. Romero, "Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses," *Comput. Electr. Eng.*, vol. 89, p. 106908, 2021, doi: 10.1016/j.compeleceng.2020.106908.

[43] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education," in *The ethics of artificial intelligence in education: Practices, challenges, and debates*, W. Holmes and K. Porayska-Pomsta, Eds., Routledge, 2022, pp. 174–202. doi: 10.4324/9780429329067-10.

[44] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," *Am. Behav. Sci.*, vol. 57, no. 10, pp. 1510–1529, 2013, doi: 10.1177/0002764213479366.