Sentiment Analysis of Student Reviews on Lecturer Performance Using a Web-Based Naive Bayes Method

Daniel Chua[1]), Hanifah Ekawati[2]), dan Wahyuni[3])
[1],[2],[3]Teknik Informatika, STMIK Widya Cipta Dharma
[1],[2],[3]Jl. M. Yamin No.25, Gn. Kelua, Samarinda, Kalimantan Timur, 75123
E-mail: [email Daniel Chua][1]), [email Hanifah Ekawati][2]), [email Wahyuni][3])

ABSTRACT

Student textual reviews provide important qualitative information for evaluating lecturer performance, but manual processing is inefficient and vulnerable to subjectivity, while the data collected can also be imbalanced across sentiment classes. This study aims to classify student reviews of lecturer performance into positive, neutral, and negative sentiments by applying Multinomial Naive Bayes and to implement the classification model in a web-based evaluation system. The research follows the CRISP-DM stages, covering business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Data were collected through questionnaires from 151 active students, generating 755 raw reviews; after manual cleaning, 413 valid reviews remained, consisting of 217 positive, 174 negative, and 22 neutral reviews. Text preprocessing included case folding, cleaning, tokenizing, slang-word normalization, stopword removal, and stemming. The data were split into 80% training and 20% testing sets. TF-IDF with an N-Gram range of (1,3) was used to transform text, and Random Over Sampling was applied only to the training data to reduce class imbalance. The final Multinomial Naive Bayes model used a Laplace Smoothing parameter of 0.1. Testing on 83 fixed test reviews produced 80.72% accuracy, 72.51% macro-average precision, 79.80% macro-average recall, and 74.89% macro-average F1-score. The model was deployed using React, Flask, and MySQL, and all tested web functions passed Black Box Testing.
Keywords: Sentiment Analysis, Multinomial Naive Bayes, Lecturer Performance Evaluation, CRISP-DM, Random Over Sampling

Analisis Sentimen Ulasan Mahasiswa terhadap Kinerja Dosen Menggunakan Metode Naive Bayes Berbasis Web

ABSTRAK

Ulasan tekstual mahasiswa menyediakan informasi kualitatif penting dalam evaluasi kinerja dosen, tetapi pengolahan secara manual kurang efisien dan rentan terhadap subjektivitas, sementara data yang terkumpul juga dapat memiliki distribusi kelas sentimen yang tidak seimbang. Penelitian ini bertujuan mengklasifikasikan ulasan mahasiswa terhadap kinerja dosen ke dalam sentimen positif, netral, dan negatif menggunakan Multinomial Naive Bayes serta mengimplementasikan model klasifikasi pada sistem evaluasi berbasis web. Tahapan penelitian mengikuti CRISP-DM yang meliputi business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Data diperoleh melalui kuesioner dari 151 mahasiswa aktif yang menghasilkan 755 ulasan mentah; setelah pembersihan manual diperoleh 413 ulasan valid, terdiri atas 217 positif, 174 negatif, dan 22 netral. Prapemrosesan meliputi case folding, cleaning, tokenizing, normalisasi slangword, stopword removal, dan stemming. Data dibagi menjadi 80% data latih dan 20% data uji. Teks ditransformasikan menggunakan TF-IDF dengan N-Gram (1,3), sedangkan Random Over Sampling diterapkan hanya pada data latih untuk mengurangi ketidakseimbangan kelas. Model akhir Multinomial Naive Bayes menggunakan parameter Laplace Smoothing 0,1. Pengujian terhadap 83 data uji tetap menghasilkan akurasi 80,72%, macro average precision 72,51%, macro average recall 79,80%, dan macro average F1-score 74,89%. Model kemudian diimplementasikan menggunakan React, Flask, dan MySQL, dan seluruh fungsi yang diuji melalui Black Box Testing berstatus lulus.
Kata Kunci: Analisis Sentimen, Multinomial Naive Bayes, Evaluasi Kinerja Dosen, CRISP-DM, Random Over Sampling

1. PENDAHULUAN

Evaluasi kinerja dosen melalui ulasan mahasiswa menjadi bagian penting dalam pemantauan mutu pembelajaran karena ulasan tekstual dapat menangkap kritik, saran, dan apresiasi yang tidak selalu terwakili oleh penilaian angka. Di STMIK Widya Cipta Dharma, pengolahan ulasan dalam jumlah besar secara manual menimbulkan kendala efisiensi dan subjektivitas. Karakter bahasa mahasiswa yang kasual, menggunakan singkatan, slang, serta struktur kalimat yang beragam juga memperumit identifikasi sentimen secara konsisten. Kondisi tersebut mendorong perlunya sistem analisis sentimen otomatis yang dapat mengelompokkan opini mahasiswa ke dalam sentimen positif, netral, dan negatif.

Permasalahan lain yang ditemukan adalah ketidakseimbangan distribusi kelas. Dari 413 ulasan valid, terdapat 217 ulasan positif, 174 ulasan negatif, dan hanya 22 ulasan netral. Ketimpangan ini berpotensi membuat model klasifikasi lebih banyak mempelajari pola kelas mayoritas dan kurang mengenali kelas minoritas. Oleh karena itu, penelitian menerapkan Random Over Sampling (ROS) pada data pelatihan agar setiap kelas

memperoleh kesempatan belajar yang lebih seimbang, sementara data pengujian tetap dipertahankan dalam kondisi asli untuk menjaga objektivitas evaluasi.

Penelitian sebelumnya menunjukkan bahwa Naive Bayes telah digunakan untuk analisis sentimen mahasiswa. Eunike (2024) menerapkan Naive Bayes untuk mengelompokkan sentimen kepuasan mahasiswa terhadap laboratorium komputer. Sasmita, Pradnyana, dan Divayana (2022) mengembangkan sistem analisis sentimen untuk evaluasi kinerja dosen dan memperoleh akurasi 90,60%, tetapi masih mencatat salah klasifikasi pada sentimen negatif akibat ketidakseimbangan data. Fatmawati, Prasetya, Prastya, dan Cipta (2025) juga menerapkan Multinomial Naive Bayes untuk evaluasi kinerja pengajaran dosen dan memperoleh akurasi 94%. Berdasarkan kajian tersebut, celah penelitian terletak pada penanganan ketidakseimbangan kelas dan integrasi model ke sistem evaluasi internal kampus.

Kebaruan penelitian ini adalah penggunaan ROS pada matriks TF-IDF data latih, pengujian kombinasi N-Gram sampai trigram, serta implementasi model Multinomial Naive Bayes ke dalam aplikasi berbasis web untuk mendukung Unit Penjaminan Mutu (UPM) dan program studi. Solusi yang dikembangkan tidak hanya menghasilkan klasifikasi sentimen, tetapi juga menyediakan dashboard statistik, peringkat kinerja, pengelolaan kamus text mining, pelatihan ulang model, dan panel evaluasi model. Dengan demikian, penelitian diarahkan untuk menghasilkan proses evaluasi ulasan dosen yang lebih cepat, terstruktur, dan dapat dipantau melalui aplikasi web.

2. RUANG LINGKUP

Penelitian mencakup pengolahan ulasan tekstual mahasiswa mengenai kinerja dosen di STMIK Widya Cipta Dharma. Data diklasifikasikan ke dalam tiga kelas sentimen, yaitu positif, netral, dan negatif. Algoritma klasifikasi yang digunakan dibatasi pada Multinomial Naive Bayes, ekstraksi fitur menggunakan TF-IDF dengan N-Gram satu sampai tiga kata, dan penanganan ketidakseimbangan kelas menggunakan Random Over Sampling pada data latih. Aplikasi dikembangkan berbasis web dan dikhususkan untuk lingkungan internal kampus.

Batasan penelitian meliputi tidak diprosesnya data audio, gambar, maupun penilaian kuantitatif berupa angka; data uji tidak dikenai ROS; aplikasi tidak diperuntukkan bagi mata kuliah dengan team teaching; dan pengujian aplikasi berfokus pada fungsi melalui Black Box Testing. Hasil yang direncanakan adalah model klasifikasi sentimen yang terukur dengan accuracy, precision, recall, dan F1-score serta aplikasi yang mampu menyajikan hasil analisis melalui dashboard interaktif.

3. BAHAN DAN METODE

3.1 Analisis Sentimen dan Prapemrosesan

Analisis sentimen merupakan proses untuk mengenali dan mengelompokkan opini dalam teks ke dalam kategori positif, negatif, atau netral (Lukmana, L, & Elisya, 2025). Karena data penelitian berbentuk ulasan bebas, teks perlu dipersiapkan sebelum klasifikasi. Tahapan prapemrosesan terdiri atas case folding, cleaning, tokenizing, normalisasi kata tidak baku menggunakan kamus slangword, stopword removal, dan stemming menggunakan Sastrawi. Prapemrosesan bertujuan mengubah teks yang tidak terstruktur menjadi bentuk yang lebih seragam dan siap diproses oleh model (Agung & Dwi, 2023). Pendekatan pengolahan teks tidak terstruktur tersebut juga merupakan bagian dari text mining (Ridwansyah, 2022; Sukriadi, Ismail, & Andzar, 2023).

3.2 TF-IDF dan Multinomial Naive Bayes

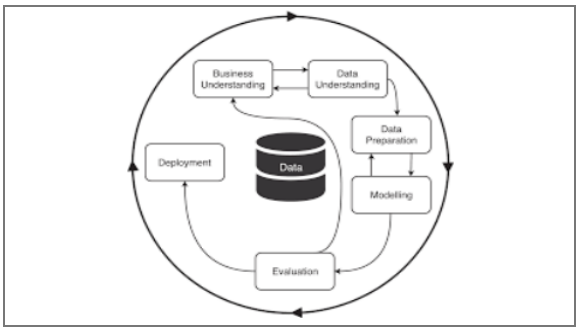
TF-IDF digunakan untuk mengubah data teks menjadi representasi numerik berdasarkan tingkat kepentingan kata pada dokumen dan korpus (Septiani & Isabela, 2023). Vectorizer dibentuk hanya dari data pelatihan menggunakan fit_transform(), sedangkan data pengujian menggunakan transform() dengan vectorizer yang sama. Konfigurasi akhir memakai N-Gram (1,3), sehingga fitur mencakup unigram, bigram, dan trigram. Multinomial Naive Bayes dipilih karena sesuai untuk klasifikasi data teks dan menentukan kelas berdasarkan distribusi fitur kata pada data berlabel (Permana, Chrisnanto, & Ashaury, 2023).

3.3 Random Over Sampling dan CRISP-DM

Random Over Sampling merupakan teknik penyeimbangan data dengan mereplikasi sampel kelas minoritas secara acak sampai distribusi kelas pada data pelatihan menjadi lebih seimbang (Alfinsyah & Hartato, 2025). Dalam penelitian ini ROS hanya diterapkan pada matriks TF-IDF data pelatihan untuk menghindari kebocoran data dan menjaga data pengujian tetap merepresentasikan data baru. Kerangka pengembangan mengikuti CRISP-DM yang terdiri atas business understanding, data understanding, data preparation, modeling, evaluation, dan deployment (Rianti, Majid, & Fauzi, 2023).

3.4 Tahapan Penelitian

Penelitian dilakukan di STMIK Widya Cipta Dharma sejak Maret 2026. Pengumpulan data utama dilakukan melalui kuesioner kepada mahasiswa aktif. Berdasarkan perhitungan Slovin dengan tingkat kesalahan 10%, kebutuhan minimal adalah 94 responden, sedangkan penelitian memperoleh 151 responden. Setiap responden memberikan lima ulasan sehingga terkumpul 755 ulasan mentah. Setelah pembersihan manual untuk menghapus ulasan kosong, hanya berisi tanda baca, atau tidak relevan, diperoleh 413 ulasan valid. Penentuan kebutuhan minimal sampel mengacu pada konsep Slovin, sedangkan pemilihan sampel diarahkan agar mewakili populasi penelitian (Majdina, Pratikno, & Tripena, 2024; Suriani, Risnita, & Jailani, 2023).



Gambar 1. Alur Penelitian Berbasis CRISP-DM
Figure 1. CRISP-DM-Based Research Flow

Sebanyak 413 ulasan dilabeli manual berdasarkan isi dan makna utama. Ulasan positif berisi apresiasi atau penilaian baik, ulasan netral bersifat informatif atau saran tanpa kecenderungan kuat, sedangkan ulasan negatif berisi kritik, keluhan, atau ketidakpuasan. Distribusi awal dataset ditunjukkan pada Tabel 1.

Distribusi dataset: Positif 217 ulasan; Negatif 174 ulasan; Netral 22 ulasan.

Data dibagi dengan proporsi 80% pelatihan dan 20% pengujian menggunakan random_state=0, menghasilkan 330 data latih dan 83 data uji. TF-IDF N-Gram (1,3) membentuk 2.991 fitur, dengan matriks pelatihan berukuran 330 × 2.991 dan matriks pengujian 83 × 2.991. ROS selanjutnya diterapkan pada data pelatihan. Pada bagian hasil rinci skripsi, distribusi data latih sebelum ROS tercatat 171 positif, 140 negatif, dan 19 netral, kemudian diseimbangkan menjadi masing-masing 171 data dengan total 513 data latih. Model dilatih menggunakan Multinomial Naive Bayes dan nilai alpha terbaik dipilih melalui pengujian parameter.

Pembentukan fitur dijaga agar tidak terjadi kebocoran data. TF-IDF dibentuk menggunakan fit_transform() pada data latih, sedangkan data uji hanya menggunakan transform() dari vectorizer yang telah terbentuk. Random Over Sampling juga hanya diterapkan pada data latih. Dengan demikian, 83 data uji tetap berada pada kondisi asli dan digunakan sebagai dasar untuk menilai kemampuan model terhadap data yang tidak dilibatkan pada proses pelatihan.

4. PEMBAHASAN

4.1 Hasil Prapemrosesan dan Pembentukan Fitur

Tahap prapemrosesan diterapkan secara berurutan pada 413 ulasan valid. Case folding menyeragamkan seluruh huruf menjadi huruf kecil. Cleaning menghapus angka dan tanda baca yang tidak dibutuhkan. Tokenizing memecah kalimat menjadi unit kata, kemudian normalisasi mengubah kata tidak baku menjadi bentuk baku berdasarkan kamus slangword. Stopword removal menghapus kata umum dengan kontribusi informasi rendah, sedangkan stemming menggunakan Sastrawi mengubah kata berimbuhan menjadi kata dasar. Rangkaian ini menghasilkan bentuk teks yang lebih seragam sehingga variasi penulisan tidak memperbesar ruang fitur secara tidak perlu. Tahapan case folding dan penyaringan kata mengacu pada prinsip prapemrosesan teks yang bertujuan menyeragamkan bentuk data sebelum klasifikasi (Wajhillah & Wibowo, 2022; Angelina, Negara, & Muhardi, 2023).

Contoh pada skripsi menunjukkan kalimat ‘UTS dan UAS ditiadakan, diganti proyek akhir’ berubah menjadi huruf kecil pada case folding, kemudian tanda baca dihapus pada cleaning. Pada tahap tokenizing dan normalisasi, bentuk ‘uts’ dan ‘uas’ dinormalisasi menjadi ‘ujian tengah semester’ dan ‘ujian akhir semester’. Contoh lain, kalimat ‘terbuka sekali diajak diskusi’ setelah stemming menjadi ‘buka sekali ajak diskusi’. Transformasi ini tidak mengubah label manual, tetapi mengubah representasi teks yang digunakan model.

Contoh hasil prapemrosesan: case folding menghasilkan “uts dan uas ditiadakan, diganti proyek akhir”; cleaning menghasilkan “tidak otoriter tidak arogan”; normalisasi mengubah singkatan menjadi “ujian tengah semester dan ujian akhir semester”; stopword removal menghasilkan “banyak berinteraksi”; stemming menghasilkan “buka sekali ajak diskusi”.

Setelah prapemrosesan, data dibagi menjadi 330 data pelatihan dan 83 data pengujian. Pembentukan TF-IDF menghasilkan 2.991 fitur untuk kedua set data. Beberapa fitur dengan rata-rata bobot yang ditampilkan dalam skripsi antara lain ‘jelas’ sebesar 0,0348, ‘tidak’ sebesar 0,0322, ‘materi’ sebesar 0,0250, ‘ajar’ sebesar 0,0220, ‘sangat’ sebesar 0,0217, dan ‘baik’ sebesar 0,0182. Keberadaan kata ‘tidak’ sebagai fitur berbobot cukup tinggi juga relevan dengan temuan analisis kesalahan, karena negasi dapat muncul pada ulasan positif maupun negatif dan belum selalu dipahami berdasarkan konteks keseluruhan kalimat.

ROS diterapkan setelah teks direpresentasikan sebagai matriks TF-IDF dan hanya pada bagian pelatihan. Dengan cara ini, sampel yang direplikasi tidak pernah masuk ke data pengujian. Pendekatan tersebut penting karena data pengujian harus tetap mempertahankan distribusi asli untuk mengukur kemampuan model menghadapi data baru. Pada hasil rinci di Bab IV, data latih sebelum ROS terdiri dari 171 positif, 140 negatif, dan 19 netral, kemudian masing-masing diseimbangkan menjadi 171 data. Kelas netral memperoleh peningkatan jumlah terbesar melalui replikasi karena merupakan kelas dengan jumlah awal paling sedikit.

4.2 Hasil Pengujian Konfigurasi Model

Pengujian dilakukan pada data uji yang tidak digunakan dalam proses pelatihan. Dua nilai random state dibandingkan pada pembagian 80:20. random_state=42 menghasilkan akurasi 75,90%, sedangkan random_state=0 menghasilkan 80,72%, sehingga random_state=0 digunakan pada pembagian data final. Nilai random_state pada ROS tetap 42 karena digunakan untuk proses yang berbeda, yaitu menjaga replikasi sampel minoritas tetap konsisten.

Pengujian Laplace Smoothing menunjukkan alpha 0,1 menghasilkan akurasi tertinggi 80,72%. Alpha 0,5 menghasilkan 78,31% dan alpha 1,0 menghasilkan

77,11%. Setelah itu, pengujian fitur menunjukkan N-Gram (1,3) menghasilkan akurasi 80,72%, lebih tinggi daripada N-Gram (1,2) sebesar 79,52%. Hasil ini menjadi dasar pemilihan alpha 0,1 dan N-Gram (1,3) pada model final.

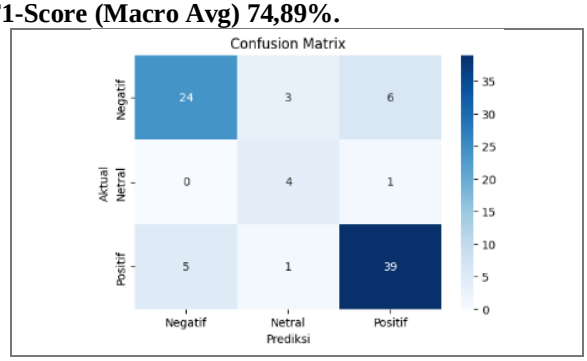
Ringkasan pengujian konfigurasi: random_state=42 menghasilkan akurasi 75,90%; random_state=0 menghasilkan 80,72%; alpha 0,5 menghasilkan 78,31%; alpha 1,0 menghasilkan 77,11%; N-Gram (1,2) dengan alpha 0,1 menghasilkan 79,52%; N-Gram (1,3) dengan alpha 0,1 menghasilkan 80,72%.

Konfigurasi akhir menggunakan random_state=0 untuk pembagian data, alpha 0,1, dan N-Gram (1,3). Kombinasi trigram membantu model menangkap hubungan antarkata secara lebih spesifik dibanding penggunaan unigram dan bigram saja, sehingga representasi fitur menjadi lebih informatif pada data ulasan mahasiswa.

4.3 Performa Model Final

Model final diuji pada 83 ulasan. Sebanyak 67 data diprediksi benar dan 16 data masih salah klasifikasi, menghasilkan accuracy 80,72%. Nilai macro average precision mencapai 72,51%, macro average recall 79,80%, dan macro average F1-score 74,89%. Nilai recall yang lebih tinggi daripada precision menunjukkan bahwa model relatif baik dalam menemukan data aktual setiap kelas, meskipun ketepatan sebagian prediksi masih perlu ditingkatkan.

Performa model final: Accuracy 80,72%; Precision (Macro Avg) 72,51%; Recall (Macro Avg) 79,80%; F1-Score (Macro Avg) 74,89%.



Gambar 2. Confusion Matrix Model Final
Figure 2. Final Model Confusion Matrix

Confusion matrix memperlihatkan bahwa 39 dari 45 ulasan positif diprediksi dengan benar; lima ulasan positif diprediksi negatif dan satu diprediksi netral. Pada kelas negatif, 24 dari 33 data diprediksi benar, sedangkan enam diprediksi positif dan tiga diprediksi netral. Pada kelas netral, empat dari lima data diprediksi benar dan satu diprediksi positif. Hasil tersebut menunjukkan bahwa model dapat mengenali mayoritas pola pada semua kelas, tetapi kelas dengan jumlah data awal yang lebih sedikit tetap memiliki keragaman bahasa yang terbatas.

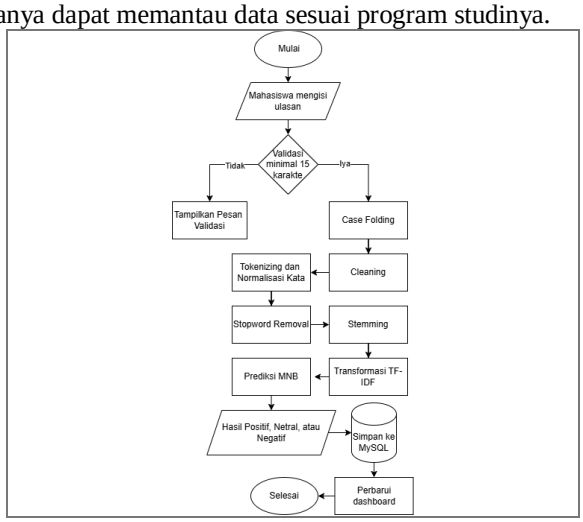
Analisis kesalahan menunjukkan beberapa sumber kekeliruan. Ulasan negatif dapat memuat kata seperti ‘baik’, ‘mudah’, atau ‘akrab’ yang sering muncul pada kelas positif. Sebaliknya, ulasan positif dengan negasi seperti ‘tidak otoriter’, ‘tidak arogan’, atau ‘tidak perlu diperbaiki lagi’ dapat dipahami model sebagai pola negatif karena Multinomial Naive Bayes mengandalkan probabilitas kemunculan fitur dan belum memahami perubahan makna akibat negasi secara mendalam. Ulasan netral ‘bisa lebih baik lagi’ juga diprediksi positif karena kata ‘baik’ memiliki asosiasi yang kuat dengan kelas positif.

Secara rinci, model mengenali 39 dari 45 ulasan positif, 24 dari 33 ulasan negatif, dan 4 dari 5 ulasan netral dengan benar. Kesalahan terutama muncul ketika kata tertentu memiliki kecenderungan sentimen yang berbeda dengan makna keseluruhan kalimat. Kondisi ini memperlihatkan keterbatasan Multinomial Naive Bayes dalam memahami negasi, kritik tidak langsung, dan konteks semantik yang lebih kompleks.

4.4 Implementasi Sistem Berbasis Web

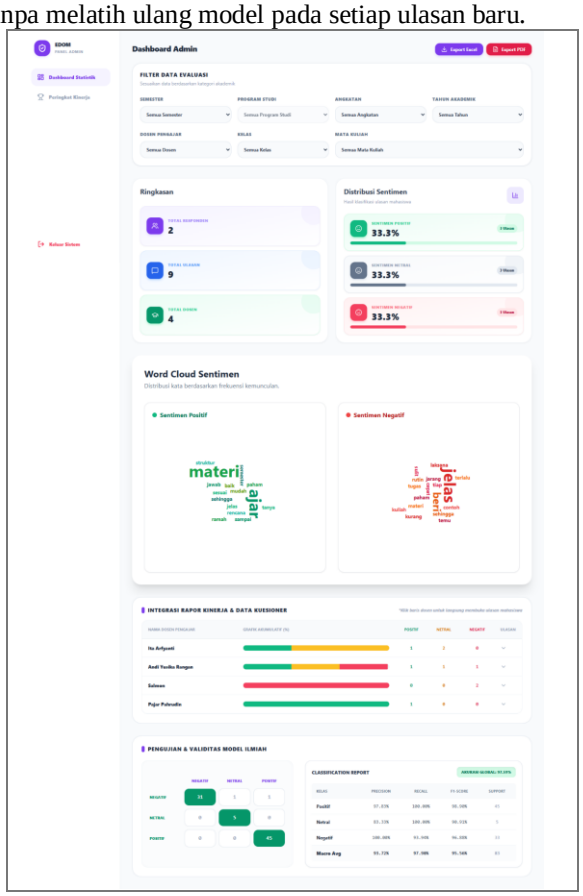
Model klasifikasi disimpan bersama TF-IDF Vectorizer menggunakan Joblib dan diintegrasikan ke sistem berbasis web. Antarmuka menggunakan React, backend menggunakan Flask, dan penyimpanan data menggunakan MySQL. Sistem menyediakan akses mahasiswa untuk login, melihat progres, dan mengisi kuesioner minimal 15 karakter. Ulasan yang valid diproses oleh model dan hasilnya disimpan ke basis data. Pemanfaatan React dan Flask sebagai teknologi antarmuka dan layanan backend sejalan dengan penggunaannya pada pengembangan aplikasi web dalam penelitian terdahulu (Iswari, 2021; Wijayanto & Susetyo, 2022).

Pada sisi pengelola, Admin UPM dapat melihat ringkasan statistik, distribusi sentimen, word cloud, integrasi rapor kinerja dan data kuesioner, serta panel evaluasi model yang menampilkan confusion matrix, classification report, accuracy, precision, recall, dan F1-score. Admin UPM juga dapat mengelola data akademik, memperbarui kamus slangword dan stopword, melakukan koreksi label, serta menjalankan retraining. Pengguna program studi memiliki hak akses yang lebih terbatas dan hanya dapat memantau data sesuai program studinya.



Gambar 3. Alur Proses Klasifikasi Sentimen
Figure 3. Sentiment Classification Process Flow

Pada alur operasional, ulasan mahasiswa yang masuk diproses dengan tahapan prapemrosesan yang sama seperti data pelatihan, kemudian ditransformasikan menggunakan TF-IDF Vectorizer yang telah disimpan. Matriks fitur hasil transformasi menjadi masukan model Multinomial Naive Bayes untuk menghasilkan label sentimen. Hasil klasifikasi disimpan ke basis data dan selanjutnya direkap pada dashboard berdasarkan tahun akademik, semester, program studi, mata kuliah, dan dosen. Pemisahan antara model, vectorizer, backend, dan antarmuka memungkinkan proses prediksi dijalankan tanpa melatih ulang model pada setiap ulasan baru.



Gambar 4. Dashboard Admin untuk Pemantauan Evaluasi

Figure 4. Admin Dashboard for Evaluation Monitoring

4.5 Pengujian Sistem dan Keterbatasan

Black Box Testing dilakukan pada fungsi utama mahasiswa dan admin tanpa memeriksa kode internal. Skenario mahasiswa mencakup login, dashboard, pengisian kuesioner, serta informasi dan panduan. Pengujian admin mencakup login, dashboard, ekspor laporan, koreksi label, pembatasan hak akses, peringkat kinerja, pengelolaan data akademik, kamus text mining, retraining model, impor data, dan ekspor peringkat. Seluruh skenario yang dicantumkan pada hasil pengujian memperoleh status lulus.

Penelitian memiliki beberapa keterbatasan. Dataset hanya berasal dari mahasiswa STMIK Widya Cipta Dharma pada periode tertentu sehingga hasil belum tentu dapat digeneralisasikan ke perguruan tinggi lain. Kelas netral hanya berjumlah 22 ulasan pada dataset awal; ROS menyeimbangkan data dengan replikasi, tetapi tidak menghasilkan variasi kalimat baru. Multinomial Naive Bayes juga belum mampu memahami konteks seperti sindiran, kritik tidak langsung, makna ganda, dan negasi secara mendalam. Selain itu, panel confusion matrix dan classification report menampilkan evaluasi final yang disimpan statis pada metrics.json, sehingga belum otomatis merekam performa setiap versi model hasil retraining. Pengujian aplikasi juga belum mencakup pengujian beban, penggunaan serentak dalam skala besar, maupun keamanan secara menyeluruh.

Ringkasan Black Box Testing: login dan dashboard mahasiswa—Lulus; pengisian kuesioner dan klasifikasi sentimen—Lulus; dashboard dan evaluasi model admin—Lulus; koreksi label dan pembatasan hak akses—Lulus; pengelolaan data akademik—Lulus; kamus text mining dan retraining—Lulus; import dan export laporan—Lulus.

Hasil pengujian fungsional menunjukkan bahwa alur utama dari input ulasan sampai penyajian hasil dapat dijalankan sesuai rancangan. Namun, status lulus pada Black Box Testing tidak dapat diartikan sebagai bukti bahwa sistem telah teruji untuk beban tinggi atau serangan keamanan, karena kedua jenis pengujian tersebut memang berada di luar ruang lingkup penelitian ini.

5. KESIMPULAN

Multinomial Naive Bayes berhasil diterapkan untuk mengklasifikasikan ulasan mahasiswa ke dalam sentimen positif, netral, dan negatif. Model final menggunakan TF-IDF N-Gram (1,3) dengan Laplace Smoothing 0,1. Pengujian terhadap 83 data uji menghasilkan 67 prediksi benar, accuracy 80,72%, macro average precision 72,51%, macro average recall 79,80%, dan macro average F1-score 74,89%. Random Over Sampling digunakan hanya pada data latih untuk mengurangi ketidakseimbangan kelas dan memberi kesempatan belajar yang lebih merata bagi kelas minoritas. Model kemudian berhasil diimplementasikan ke aplikasi web menggunakan React, Flask, dan MySQL. Berdasarkan Black Box Testing, seluruh fungsi yang diuji berjalan sesuai hasil yang diharapkan.

6. SARAN

Penelitian selanjutnya perlu menambah jumlah dan keragaman ulasan, khususnya pada kelas netral, agar representasi pola bahasa menjadi lebih kuat. Kamus normalisasi slangword dan stopword perlu diperbarui secara berkala, kemudian model hasil retraining perlu diuji menggunakan dataset uji tetap untuk memastikan performanya tidak menurun. Sistem juga dapat dikembangkan dengan mekanisme pencatatan versi model, pemulihan model sebelumnya, validasi agar tidak terjadi tumpang tindih data latih dan data uji, serta pengujian lebih lanjut terhadap beban, keamanan, dan penggunaan bersamaan oleh banyak pengguna.

7. REFERENSI

Agung, A. A., & Dwi, I. D. (2023). Data preprocessing pola pada penilaian mahasiswa Program Profesi Guru. JINACS: Journal of Informatics and Computer Science, 5(1), 97.

Alfinsyah, D. R., & Hartato. (2025). Evaluating the impact of Random Over Sampling on IndoBERT performance for Indonesian sentiment analysis. JAIC, 9, 3272.

Eunike. (2024). Analisis sentimen kepuasan mahasiswa terhadap laboratorium komputer. Wicida Repository.

Fatmawati, S., Prasetya, M. A., Prastya, S. E., & Cipta, S. P. (2025). Penerapan Naïve Bayes dalam analisis sentimen untuk evaluasi kinerja pengajaran dosen. Jurnal Saintekom: Sains, Teknologi, Komputer dan Manajemen.

Lukmana, L. O., L, D. R., & Elisya, P. (2025). Analisis sentimen terhadap ulasan pengguna aplikasi Threads Instagram di Playstore menggunakan algoritma Naive Bayes. JITET (Jurnal Informatika dan Teknik Elektro Terapan), 498.

Permana, H., Chrisnanto, Y. H., & Ashaury, H. (2023). Analisis sentimen terhadap bakal calon presiden 2024 dengan. JATI (Jurnal Mahasiswa Teknik Informatika), 3257–3264.

Pratama, S. D., Lasimin, & Dadaprawira, M. N. (2023). Pengujian Black Box Testing pada aplikasi Edu Digital berbasis website menggunakan metode equivalence dan boundary value. Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD, 563.

Rianti, A., Majid, N. W., & Fauzi, A. (2023). CRISP-DM: Metodologi proyek data science. SENATIB, 110.

Sasmita, A., Pradnyana, G. A., & Divayana, D. G. (2022). Sistem analisis sentimen untuk evaluasi kinerja dosen dengan metode Naïve Bayes. Jurnal Sains dan Teknologi.

Septiani, D., & Isabela, I. (2023). Analisis Term Frequency Inverse Document Frequency (TF-IDF) dalam temu kembali informasi pada dokumen teks. Analisis Term Frequency Inverse Document Frequency (TF-IDF) dalam Temu Kembali Informasi pada Dokumen Teks, 82.

Angelina, S. J., Negara, A. B., & Muhardi, H. (2023). Analisis pengaruh penerapan stopword removal pada performa. Jurnal Aplikasi dan Riset Informatika, 167–168.

Iswari, L. (2021). Penerapan React JS pada pengembangan frontend aplikasi startup Ubaform. Automata, 2.

Majdina, N. I., Pratikno, B., & Tripena, A. (2024). Penentuan ukuran sampel menggunakan rumus Bernoulli dan Slovin: Konsep dan aplikasinya. JMP: Jurnal Ilmiah Matematika dan Pendidikan Matematika, 73–84. https://doi.org/10.20884/1.JMP.2024.16.1.11230

Ridwansyah, T. (2022). Implementasi text mining terhadap analisis sentimen masyarakat dunia di Twitter terhadap Kota Medan menggunakan K-Fold Cross Validation dan Naïve Bayes Classifier. KLIK: Kajian Ilmiah Informatika dan Komputer, 179.

Sukriadi, Ismail, & Andzar, A. M. (2023). Penerapan text mining dalam klasifikasi judul skripsi yang diusulkan mahasiswa menggunakan metode Naïve Bayes. Jurnal Ilmiah Sistem Informasi dan Teknik Informatika (JISTI), 186.

Suriani, N., Risnita, & Jailani, M. S. (2023). Konsep populasi dan sampling serta pemilihan partisipan ditinjau dari penelitian ilmiah pendidikan. IHSAN: Jurnal Pendidikan Islam, 24–36.

Wajhillah, R., & Wibowo, A. (2022). Information retrieval pemetaan peta jalan penelitian. Jurnal LARIK, 52–53.

Wijayanto, C., & Susetyo, Y. A. (2022). Implementasi Flask Framework pada pembangunan aplikasi Sistem Informasi Helpdesk (SIH). JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 7, 858–868. https://doi.org/10.29100/JIPI.V7I3.3161