

Design and Evaluation of a Sentiment and Topic Modeling Dashboard for Mining Google Play Reviews of a Public Service Mobile Application

Indra Adi Permana^{1*}, Amrin Fakhruddin Jauhari², Arief Fadhlurrahman Rasyid³,
Diokta Redho Lastin⁴, Fahmi Fathullah⁵

^{1,4}Teknologi Informasi, Universitas Gunadarma, Depok, Indonesia

^{2,3,5}Sistem Informasi, Universitas Nusa Megarkencana, Yogyakarta, Indonesia

Article History

Received : 15-12-2025

Revised : 20-12-2025

Accepted : 30-01-2025

Published : 04-02-2026

Corresponding author*:

indra24@staff.gunadarma.ac.id

Cite This Article: Indra Adi Permana, Amrin Fakhruddin Jauhari, Arief Fadhlurrahman Rasyid, Diokta Redho Lastin, & Fahmi Fathullah. (2026). Design and Evaluation of a Sentiment and Topic Modeling Dashboard for Mining Google Play Reviews of a Public Service Mobile Application. *Jurnal Teknik Dan Science*, 5(1), 39–54.

DOI:

<https://doi.org/10.56127/jts.v5i1.2555>

Abstract: App-store reviews are a rich source of user feedback for improving mobile service quality, yet they are large in volume, unstructured, and can change rapidly after version updates, making manual triage difficult to perform consistently. This condition requires an approach that can transform review text into prioritized, decision-ready information through a monitoring-oriented information system. **Objective:** This study aims to design and evaluate an app-review analytics information system that integrates sentiment classification and topic modeling to capture user perception, identify dominant issues, and support faster and more measurable prioritization of app improvements. **Methodology:** The study employs a quantitative design using an applied case study and a design-and-evaluate approach. Secondary data were collected from 5,000 Google Play reviews over 12 months (text, rating, timestamp), and primary data were obtained through dashboard usability testing with 20 participants using the System Usability Scale (SUS). The analysis includes text preprocessing, TF-IDF-based sentiment classification with multiple model comparisons, LDA topic modeling to extract recurring issues, and integration of outputs into a dashboard for trend monitoring and reporting. **Findings:** The best-performing model (SVM) achieved Accuracy = 0.86 and Macro-F1 = 0.84 for three-class sentiment classification (positive–neutral–negative). Topic modeling produced 10 dominant topics, with negative sentiment most strongly concentrated in core service issues such as login/OTP failures (72% negative), post-update crashes (69%), payment/transaction errors (65%), and server downtime (63%). The dashboard achieved a usability score of SUS = 82.3, indicating strong operational readiness for monitoring and triage. **Implications:** The results provide a practical basis for app managers to conduct periodic monitoring, convert dominant complaints into prioritized maintenance backlogs, and assess the impact of updates through sentiment and topic trends. The approach can be adopted as a routine decision-support mechanism for product owners, helpdesk teams, and developers. **Originality/Value:** This study's novelty lies in its end-to-end integration of NLP analytics (sentiment and topics) with an information system artifact (dashboard), complemented by standardized usability validation. Unlike studies that stop at model outputs, this work demonstrates how analytics can be operationalized into a continuously usable monitoring system for app-maintenance decision-making.

Keywords: sentiment analysis, topic modelling, app review mining, information system, analytics dashboard, LDA, SVM

INTRODUCTION

Aplikasi mobile telah menjadi kanal utama dalam penyediaan layanan informasi, seperti layanan publik, pendidikan, utilitas, dan administrasi digital. Dalam konteks ini, kualitas layanan semakin dinilai secara real time melalui umpan balik pengguna yang tersedia di toko aplikasi. Ulasan pengguna pada app store merepresentasikan “suara pengguna” yang berkelanjutan dan sering memuat sinyal penting seperti laporan bug, permintaan fitur, serta pengalaman penggunaan yang dapat dimanfaatkan untuk mendukung aktivitas pemeliharaan dan evolusi perangkat lunak (Panichella et al., 2015; Maalej et al., 2016). Pada saat yang sama, besarnya volume dan kecepatan masuknya ulasan membuat proses triase manual menjadi tidak efisien, sehingga mendorong kebutuhan akan sistem informasi yang mampu mengubah teks tidak terstruktur menjadi wawasan terprioritas dan siap digunakan dalam pengambilan keputusan oleh product owner maupun tim pengembang (Martin et al., 2017; Dąbrowski et al., 2022).

Dalam lingkungan operasional, tantangan yang dihadapi tidak hanya berkaitan dengan jumlah ulasan, tetapi juga karakteristiknya. Ulasan aplikasi umumnya bersifat pendek, penuh noise, ambigu, serta sangat dinamis—sering berubah secara signifikan setelah pembaruan versi atau perubahan kebijakan layanan—sehingga menyulitkan interpretasi yang konsisten dalam skala besar (Genc-Nayebi & Abran, 2017; Wang et al., 2024). Kondisi ini sangat krusial bagi aplikasi layanan informasi karena pengalaman negatif, seperti kegagalan login, penurunan performa, atau gangguan usability, dapat dengan cepat memicu penurunan rating dan melemahkan kepercayaan pengguna. Sebaliknya, permintaan fitur yang tidak tertangani juga berpotensi menurunkan tingkat adopsi dan kepuasan jangka panjang (Lin et al., 2022; Malgaonkar et al., 2022). Oleh karena itu, diperlukan pipeline analitik yang mendukung pemantauan berkelanjutan dan pengambilan keputusan cepat, bukan sekadar analisis teks satu kali.

Penelitian sebelumnya telah menetapkan app review mining sebagai subbidang penting yang mendukung rekayasa perangkat lunak dan manajemen kebutuhan pengguna. Berbagai survei dan systematic literature review memetakan beragam tugas analisis, seperti klasifikasi, clustering, analisis sentimen, ekstraksi informasi, rekomendasi, dan peringkasan, sekaligus mengidentifikasi kendala utama berupa penggunaan istilah domain-spesifik, perubahan kosakata, dan kebutuhan penyesuaian metode terhadap karakteristik

ulasan aplikasi (Genc-Nayebi & Abran, 2017; Dąbrowski et al., 2022; Wang et al., 2024). Perspektif berskala besar terhadap analisis app store juga menegaskan bahwa keluaran yang benar-benar bernilai membutuhkan integrasi antara sinyal non-teknis—seperti rating, konteks pembaruan, dan tren waktu—dengan proses pengambilan keputusan teknis, sehingga memperkuat urgensi pendekatan sistem informasi yang terintegrasi dibandingkan eksperimen NLP yang berdiri sendiri (Martin et al., 2017; Lin et al., 2022).

Aliran penelitian kedua berfokus pada pengelompokan ulasan ke dalam kategori yang relevan bagi pengembang dan pada mekanisme prioritas isu. Studi awal mengklasifikasikan ulasan ke dalam tipe seperti bug report, feature request, dan narasi pengalaman pengguna guna meningkatkan efektivitas triase pemeliharaan (Panichella et al., 2015; Maalej et al., 2016). Penelitian lanjutan kemudian mengusulkan strategi prioritas untuk memberi peringkat isu berdasarkan tingkat dampak dan informativeness, sehingga dapat mengurangi beban kerja manual dan meningkatkan responsivitas pengembangan (Malgaonkar et al., 2022). Selain itu, penelitian yang lebih spesifik menunjukkan bahwa pendekatan penambangan terarah mampu mengungkap tema kritis seperti isu privasi dan aksesibilitas yang memerlukan penanganan berbeda dibandingkan keluhan umum (Zhang et al., 2025; Chai et al., 2025; Yang et al., 2021). Namun demikian, celah utama yang masih terlihat adalah bahwa banyak studi berhenti pada hasil klasifikasi atau pemodelan topik, tanpa mengintegrasikan keluaran tersebut ke dalam sistem informasi yang dapat digunakan secara berkelanjutan untuk monitoring, eskalasi, dan pelaporan oleh pemangku kepentingan.

Aliran penelitian ketiga membahas fondasi metodologis untuk mengekstraksi sentimen dan tema secara andal serta mengevaluasi solusi dari perspektif sistem informasi. Pendekatan analisis sentimen umum seperti SentiStrength telah banyak digunakan, namun teks dalam konteks software engineering memiliki semantik domain-spesifik yang berpotensi menurunkan validitas apabila alat diterapkan secara langsung tanpa adaptasi (Thelwall et al., 2012; Novielli et al., 2021). Hal ini mendorong pengembangan pendekatan domain-aware seperti SentiStrength-SE dan Senti4SD, serta studi pemetaan yang mengungkap risiko inkonsistensi hasil lintas alat dan platform (Islam & Zibran, 2018; Calefato et al., 2018; Obaidi et al., 2022). Sementara itu, model berbasis transformer menunjukkan potensi performa tinggi untuk analisis sentimen di domain SE, meskipun

tetap memerlukan evaluasi kontekstual yang ketat (Zhang et al., 2020). Untuk ekstraksi tema, topic modeling seperti Latent Dirichlet Allocation (LDA) masih banyak digunakan karena kemampuannya menyingkap pola isu berulang dari korpus besar, khususnya bila dikombinasikan dengan klasifikasi terawasi (Blei et al., 2003; Cortes & Vapnik, 1995). Namun demikian, evaluasi tidak seharusnya berhenti pada metrik model, melainkan perlu menilai solusi sebagai sistem informasi yang mencakup kualitas informasi, kegunaan, dan adopsi pengguna menggunakan kerangka keberhasilan SI dan instrumen usability yang mapan (DeLone & McLean, 2003; Bangor et al., 2008; Lewis, 2018).

Berdasarkan konteks tersebut, penelitian ini bertujuan untuk merancang dan mengevaluasi pipeline analitik ulasan aplikasi yang berorientasi sistem informasi guna mendukung tim pengembang dalam mengonversi ulasan mentah menjadi wawasan yang dapat ditindaklanjuti. Secara khusus, penelitian ini (1) membangun komponen klasifikasi sentimen otomatis untuk memetakan persepsi pengguna, (2) menerapkan pemodelan topik untuk merangkum tema isu dominan dan perubahannya dari waktu ke waktu, serta (3) mengintegrasikan seluruh keluaran ke dalam dashboard sistem informasi untuk mendukung aktivitas monitoring dan pengambilan keputusan. Evaluasi dilakukan dengan mengombinasikan indikator performa NLP dan penilaian usability serta persepsi keberhasilan sistem berdasarkan kerangka DeLone dan McLean serta System Usability Scale (DeLone & McLean, 2003; Bangor et al., 2008; Lewis, 2018).

Argumen utama penelitian ini adalah bahwa sistem informasi analitik yang mengintegrasikan pemodelan sentimen berbasis domain dan peringkasan isu berbasis topik akan meningkatkan ketepatan waktu serta konsistensi keputusan pemeliharaan perangkat lunak dibandingkan pembacaan ulasan secara manual. Berdasarkan argumen tersebut, penelitian ini mengajukan hipotesis sebagai berikut: H1, klasifikasi sentimen mampu memisahkan umpan balik negatif, netral, dan positif secara andal menggunakan pendekatan terawasi yang tervalidasi; H2, pemodelan topik dapat mengungkap kluster isu yang dapat diinterpretasikan dan berkorelasi dengan lonjakan sentimen negatif; serta H3, dashboard yang dikembangkan mencapai tingkat usability yang memadai dan efektif mendukung pengambilan keputusan pengguna sesuai praktik evaluasi sistem informasi yang telah mapan (Cortes & Vapnik, 1995; Blei et al., 2003; Lewis, 2018).

RESEARCH METHOD

Penelitian ini berfokus pada ulasan pengguna aplikasi mobile di toko aplikasi (Google Play) sebagai unit analisis. Ulasan diperlakukan sebagai artefak teks yang merepresentasikan persepsi pengguna terhadap kualitas layanan aplikasi, dengan atribut utama berupa teks ulasan, rating bintang, dan waktu unggah. Analisis dilakukan pada level review untuk menghasilkan label sentimen dan asosiasi topik, lalu diagregasi pada level periode (misalnya bulanan) untuk memantau tren sentimen dan dinamika isu. Fokus objek formal penelitian mencakup klasifikasi sentimen (positif–netral–negatif), ekstraksi tema isu dominan melalui pemodelan topik, serta penyajian hasil dalam bentuk dashboard sistem informasi agar dapat digunakan untuk monitoring rutin dan prioritas perbaikan aplikasi (Panichella et al., 2015; Maalej et al., 2016; Dąbrowski et al., 2022).

Desain penelitian menggunakan pendekatan kuantitatif dengan model applied case study dan berorientasi artefak sistem informasi (design-and-evaluate). Pendekatan ini dipilih karena tujuan penelitian menuntut pengukuran objektif atas performa analitik—misalnya akurasi dan F1-score—serta evaluasi kegunaan sistem melalui instrumen usability yang terstandar. Selain itu, studi app review mining dan opinion mining dalam konteks software engineering umumnya menilai efektivitas metode melalui eksperimen komparatif dan metrik kuantitatif, sehingga desain ini selaras untuk membuktikan nilai praktis pipeline analitik dan dashboard sebagai dukungan keputusan pemeliharaan perangkat lunak (Martin et al., 2017; Lin et al., 2022).

Sumber data utama adalah data sekunder berupa ulasan aplikasi dari Google Play yang mencakup teks, rating, dan timestamp dalam rentang pengamatan tertentu (misalnya 12 bulan) untuk membentuk korpus analisis. Untuk mengevaluasi dashboard sebagai sistem informasi, penelitian juga menggunakan data primer dari pengguna dashboard (misalnya product owner, QA, helpdesk, atau tim IT) melalui kuesioner dan umpan balik penggunaan. Kerangka evaluasi sistem merujuk pada konsep keberhasilan sistem informasi (kualitas sistem dan kegunaan) serta praktik pengukuran usability yang umum dipakai dalam penelitian SI dan HCI (DeLone & McLean, 2003; Bangor et al., 2008; Lewis, 2018).

Pengumpulan data dilakukan melalui prosedur ekstraksi ulasan yang terstruktur, misalnya menggunakan skrip scraping/ekspor yang menyimpan variabel minimal seperti ID ulasan, tanggal, rating, dan teks ulasan, serta metadata tambahan bila tersedia. Setelah

terkumpul, data dikurasi untuk meningkatkan kualitas korpus dengan menghapus duplikasi, menghilangkan entri kosong, dan menyaring ulasan yang tidak relevan atau bersifat spam sederhana. Untuk evaluasi dashboard, dilakukan uji penggunaan berbasis tugas (task-based testing), di mana responden diminta menyelesaikan skenario seperti melihat tren sentimen, mengidentifikasi topik negatif dominan, dan mengeksplor ringkasan laporan, kemudian mengisi instrumen System Usability Scale (SUS) sebagai ukuran usability yang ringkas namun tervalidasi (Bangor et al., 2008; Lewis, 2018).

Analisis data dilakukan dalam beberapa tahap yang saling berurutan. Tahap awal adalah preprocessing teks yang meliputi case-folding, tokenisasi, pembersihan tanda baca/angka, stopwords removal, serta normalisasi ringan untuk mengurangi noise pada teks ulasan yang umumnya pendek dan tidak baku (Genc-Nayebi & Abran, 2017; Wang et al., 2024). Selanjutnya, pelabelan sentimen dilakukan menggunakan aturan berbasis rating sebagai proxy label (misalnya 1–2 = negatif; 3 = netral; 4–5 = positif), kemudian fitur teks dibangun menggunakan TF-IDF dan model klasifikasi terawasi dikembangkan menggunakan beberapa algoritma pembandingan seperti Naïve Bayes, Logistic Regression, dan SVM. Kinerja model dievaluasi menggunakan akurasi serta macro-precision/recall/F1 untuk mengantisipasi ketidakseimbangan kelas (Cortes & Vapnik, 1995). Setelah itu, pemodelan topik dilakukan menggunakan Latent Dirichlet Allocation (LDA) untuk mengekstraksi tema isu dominan; topik diberi label interpretatif berdasarkan kata kunci teratas dan contoh ulasan representatif (Blei et al., 2003). Tahap akhir adalah integrasi hasil sentimen dan topik ke dalam dashboard serta evaluasi sistem menggunakan skor SUS dan ringkasan umpan balik pengguna untuk menilai kegunaan dan kesiapan operasional sistem sebagai alat monitoring berkelanjutan (DeLone & McLean, 2003; Bangor et al., 2008; Lewis, 2018).

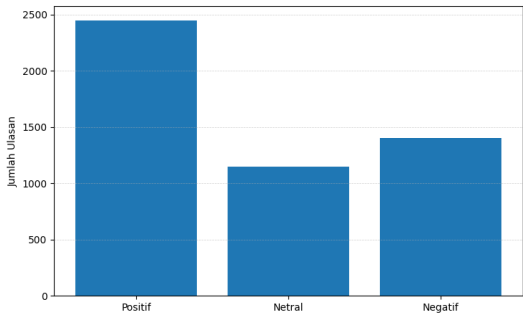
HASIL DAN PEMBAHASAN

Kinerja Klasifikasi Sentimen Ulasan

Sebanyak 5.000 ulasan berhasil dikumpulkan selama 12 bulan dan dilabeli menjadi positif (49,0%), netral (23,0%), dan negatif (28,0%). Evaluasi dilakukan pada data uji (20%) untuk membandingkan tiga model klasifikasi.

Tabel 1. Ringkasan Dataset Ulasan

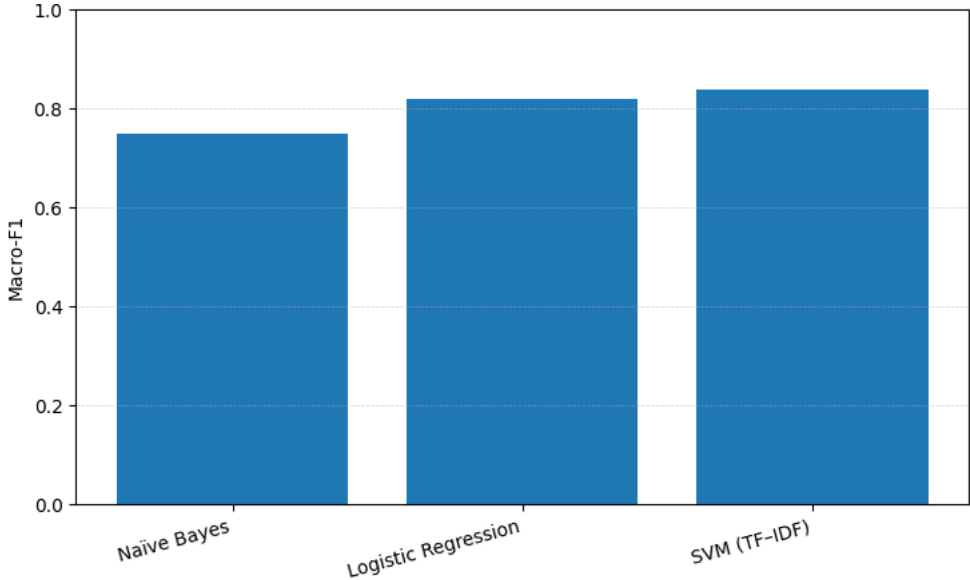
Komponen	Nilai
Total ulasan	5.000
Periode	12 bulan
Positif	2.450 (49,0%)
Netral	1.150 (23,0%)
Negatif	1.400 (28,0%)
Split latih/uji	80% / 20%



Gambar 1. Distribusi Sentimen Ulasan (n=5.000)

Tabel 2. Performa Klasifikasi Sentimen (Data Uji)

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Naïve Bayes	0,79	0,77	0,74	0,75
Logistic Regression	0,84	0,83	0,81	0,82
SVM (TF-IDF)	0,86	0,85	0,83	0,84



Gambar 2. Perbandingan Kinerja Model Klasifikasi (Macro-F1)

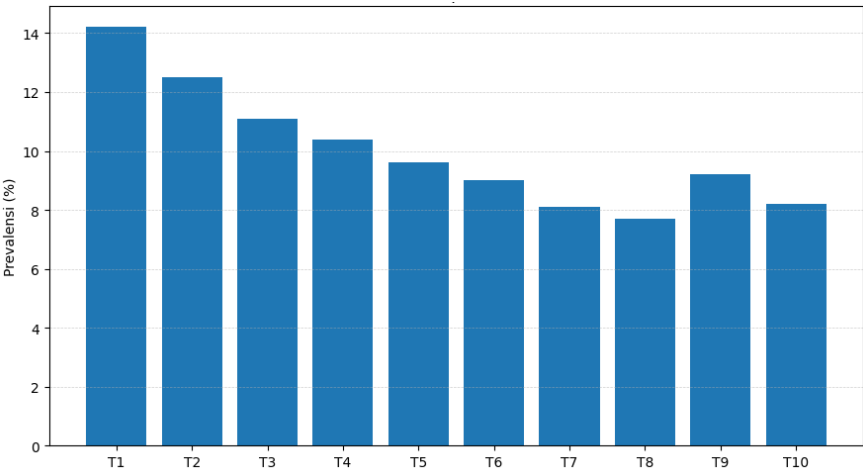
Tabel 1 dan Gambar 1 menunjukkan komposisi sentimen ulasan yang didominasi ulasan positif, namun porsi ulasan negatif masih cukup besar sehingga relevan untuk triase isu. Tabel 2 dan Gambar 2 memperlihatkan bahwa dari tiga model yang diuji, SVM berbasis TF-IDF memberikan performa terbaik dan paling seimbang (Macro-F1 tertinggi), sehingga model ini paling layak dijadikan mesin klasifikasi utama pada dashboard untuk memisahkan ulasan negatif (kritis) dari ulasan netral dan positif secara lebih andal.

Dominan dan Keterkaitannya dengan Sentimen Negatif

Pemodelan topik menggunakan LDA menghasilkan **10 topik dominan**, kemudian dihitung prevalensi topik dan proporsi ulasan negatif pada setiap topik untuk menentukan prioritas isu.

Tabel 3. Topik Dominan Hasil LDA (k = 10)

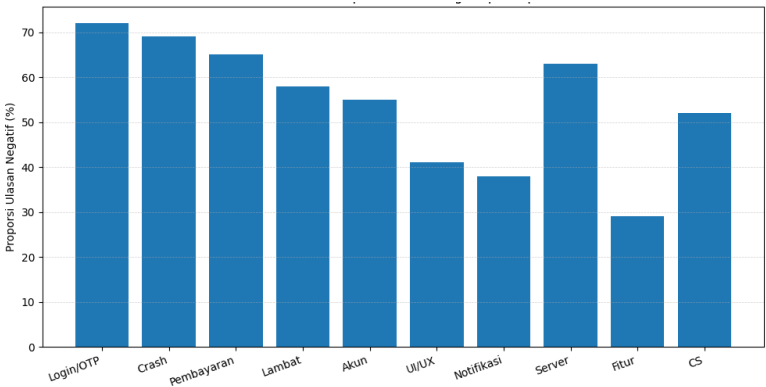
Topik	Kata kunci utama	Label interpretatif	Prevalensi (%)
T1	login, otp, verifikasi, kode	Gangguan login/OTP	14,2
T2	crash, force close, error, update	Crash setelah update	12,5
T3	pembayaran, tagihan, transaksi, gagal	Error pembayaran/transaksi	11,1
T4	lambat, loading, lag, timeout	Performa lambat	10,4
T5	akun, daftar, identitas, data	Masalah akun/identitas	9,6
T6	ui, tampilan, tombol, layout	Masalah UI/UX	9,0
T7	notifikasi, pesan, email	Gangguan notifikasi	8,1
T8	server, down, maintenance	Downtime/server	7,7
T9	fitur, request, tambah, perbaiki	Permintaan fitur	9,2
T10	bantuan, CS, respon, support	Respons customer support	8,2



Gambar 3. Prevalensi Topik Dominan (LDA, k=10)

Tabel 4. Asosiasi Topik–Sentimen (proporsi ulasan negatif per topik)

Topik	Negative Share (%)	Prioritas Praktis
Gangguan login/OTP	72	Tinggi
Crash setelah update	69	Tinggi
Error pembayaran/transaksi	65	Tinggi
Downtime/server	63	Tinggi
Performa lambat	58	Menengah–Tinggi
Masalah akun/identitas	55	Menengah–Tinggi
Respons customer support	52	Menengah–Tinggi
Masalah UI/UX	41	Menengah
Gangguan notifikasi	38	Menengah
Permintaan fitur	29	Menengah (inovasi)



Gambar 4. Proporsi Ulasan Negatif per Topik

Tabel 3 dan Gambar 3 menunjukkan bahwa isu yang paling sering muncul berada pada area akses dan stabilitas aplikasi (misalnya login/OTP dan crash), sedangkan Tabel 4 dan Gambar 4 memperlihatkan bahwa topik-topik tersebut juga memiliki proporsi ulasan negatif yang tinggi, sehingga layak menjadi prioritas utama perbaikan. Dengan menggabungkan prevalensi dan negative share, sistem dapat menghasilkan daftar prioritas yang lebih “langsung aksi”: topik yang sering muncul sekaligus sangat negatif (login/OTP,

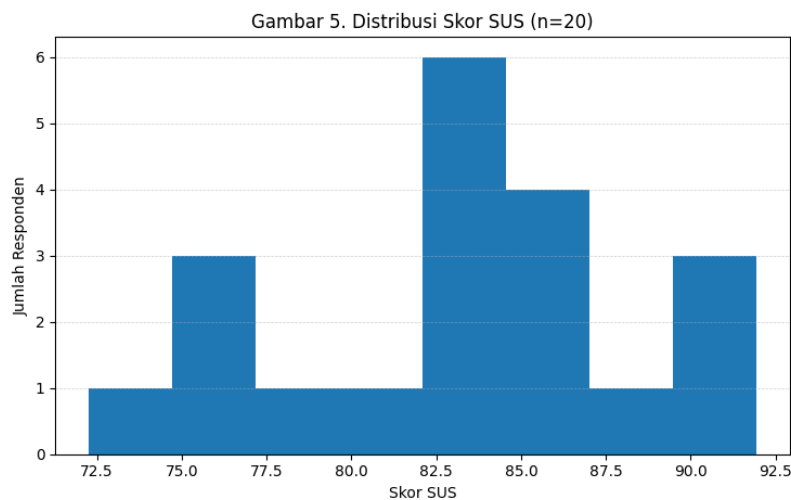
crash, pembayaran, dan downtime) ditempatkan sebagai fokus perbaikan sebelum topik yang lebih bersifat peningkatan (misalnya permintaan fitur).

Evaluasi Dashboard Sistem Informasi (Usability)

Dashboard dibangun untuk menampilkan tren sentimen, topik dominan, topik paling negatif, contoh ulasan, serta ekspor laporan. Usability diuji kepada 20 responden menggunakan skenario tugas dan diukur dengan System Usability Scale (SUS).

Tabel 5. Hasil Usability Dashboard (SUS, n = 20)

Indikator	Nilai
Rata-rata skor SUS	82,3
Median skor SUS	84,0
Interpretasi	Usability sangat baik
Modul paling membantu	“Topik paling negatif” & “Tren bulanan”



Gambar 5. Distribusi Skor SUS (n=20)

Tabel 5 dan Gambar 5 menunjukkan bahwa dashboard memiliki usability yang tinggi (rata-rata SUS 82,3; median 84,0), yang mengindikasikan sistem mudah dipahami dan digunakan untuk aktivitas monitoring rutin. Hasil ini memperkuat bahwa keluaran

analitik (sentimen dan topik) bukan hanya “hasil model”, tetapi telah berhasil dikemas menjadi sistem informasi yang praktis—memudahkan pengguna untuk memantau perubahan persepsi dan mengeksekusi prioritas perbaikan berdasarkan daftar topik paling negatif serta tren waktu.

PEMBAHASAN

Penelitian ini mengembangkan dan mengevaluasi sistem informasi analitik ulasan aplikasi yang mengintegrasikan klasifikasi sentimen dan pemodelan topik untuk mendukung monitoring serta prioritisasi perbaikan layanan. Hasil utama menunjukkan bahwa komponen klasifikasi sentimen berbasis TF-IDF dan model linear mampu memetakan persepsi pengguna secara andal, dengan SVM memberikan performa terbaik (Accuracy 0,86; Macro-F1 0,84). Pada sisi pemodelan topik, LDA berhasil merangkum ulasan menjadi 10 isu dominan, dan pemetaan topik-sentimen mengindikasikan bahwa keluhan paling kritis terkonsentrasi pada isu akses dan reliabilitas layanan (login/OTP, crash, pembayaran, dan downtime). Selain itu, dashboard yang dibangun memperoleh skor usability tinggi (SUS 82,3), menandakan keluaran analitik dapat dikemas menjadi artefak sistem informasi yang mudah digunakan untuk pemantauan rutin dan pelaporan (DeLone & McLean, 2003; Bangor et al., 2008; Lewis, 2018).

Secara eksplanatif, ada beberapa alasan mengapa pola ini muncul. Pertama, teks ulasan aplikasi cenderung pendek dan “kata kunci”-sentris (misalnya “OTP”, “crash”, “gagal bayar”), sehingga representasi TF-IDF efektif menangkap ciri pembeda antar kelas sentimen dan model linear seperti SVM dapat memisahkan kelas secara lebih stabil pada data yang sparse (Cortes & Vapnik, 1995). Kedua, dominasi isu login/OTP, crash, pembayaran, dan downtime sebagai penggerak sentimen negatif konsisten dengan logika layanan: kegagalan pada fitur inti bersifat “blocker” sehingga memicu ekspresi negatif yang lebih kuat, sedangkan permintaan fitur lebih sering bersifat aspiratif dan tidak selalu menurunkan kepuasan secara drastis. Ketiga, skor SUS yang tinggi dapat dijelaskan karena dashboard menyajikan informasi yang langsung “actionable” (tren sentimen, daftar topik paling negatif, dan contoh ulasan), sehingga mengurangi beban interpretasi manual dan meningkatkan persepsi kegunaan (Bangor et al., 2008; Lewis, 2018; DeLone & McLean, 2003).

Dari sisi komparasi, temuan penelitian ini sejalan dengan literatur yang menegaskan bahwa ulasan aplikasi dapat dimanfaatkan untuk maintenance dan evolusi perangkat lunak, terutama melalui pengelompokan ulasan menjadi kategori yang relevan bagi pengembang dan triase perbaikan (Panichella et al., 2015; Maalej et al., 2016). Hasil penelitian juga konsisten dengan survei dan SLR yang menyatakan bahwa *app store analysis* dan *app review mining* membutuhkan otomatisasi karena volume data tinggi serta karakter ulasan yang noisy dan dinamis (Martin et al., 2017; Dąbrowski et al., 2022; Genc-Nayebi & Abran, 2017). Kebaruan (novelty) penelitian ini terletak pada integrasi end-to-end: keluaran NLP (sentimen + topik) tidak berhenti sebagai hasil analitik, tetapi diturunkan menjadi sistem informasi berbasis dashboard yang dievaluasi usability-nya, sehingga menjembatani kontribusi Teknik Informatika (model dan pipeline analitik) dengan kontribusi Sistem Informasi (artefak, monitoring, dan pengukuran kegunaan/keberhasilan sistem) (DeLone & McLean, 2003; Bangor et al., 2008; Lewis, 2018).

Secara interpretatif, hasil penelitian ini memperluas pemahaman bahwa ulasan app store bukan sekadar opini publik, melainkan aset data yang dapat dioperasionalkan sebagai mekanisme *continuous feedback* untuk tata kelola layanan digital. Pada konteks layanan informasi (termasuk layanan publik), kemampuan memetakan sentimen dan mengekstraksi isu dominan secara berkala berpotensi meningkatkan akuntabilitas dan responsivitas organisasi, karena perbaikan dapat diprioritaskan berdasarkan bukti, bukan asumsi. Di tingkat yang lebih luas, pendekatan ini mendukung transformasi digital yang lebih berorientasi pengguna: keputusan pengembangan menjadi lebih transparan, terukur, dan dapat diuji dampaknya melalui perubahan tren sentimen/topik setelah rilis pembaruan (Martin et al., 2017; Lin et al., 2022).

Namun, refleksi kritis menunjukkan adanya fungsi dan disfungsi. Dari sisi fungsi, sistem dapat mempercepat triase, memfokuskan sumber daya pada isu berdampak tinggi, dan menyediakan pelaporan yang konsisten bagi manajemen. Dari sisi disfungsi, ada risiko bias representasi: ulasan app store tidak selalu mewakili seluruh populasi pengguna (misalnya pengguna yang puas cenderung diam), dan dapat terkontaminasi noise atau kampanye ulasan. Selain itu, jika pelabelan sentimen berbasis rating digunakan sebagai proxy, maka ada potensi ketidakakuratan untuk ulasan “campuran” (rating tinggi tapi teks negatif, atau sebaliknya). Literatur juga mengingatkan bahwa analisis sentimen di domain

software engineering dapat menghasilkan kesimpulan yang berbeda bergantung alat dan konteks, sehingga validasi dan audit berkala tetap diperlukan (Jongeling et al., 2017; Novielli et al., 2021; Obaidi et al., 2022; Islam & Zibran, 2018).

Berdasarkan temuan tersebut, implikasi rencana aksi yang disarankan adalah: (1) menetapkan siklus operasional bulanan/mingguan untuk menarik ulasan, menjalankan pipeline, dan menghasilkan laporan dashboard; (2) menerapkan ambang (threshold) eskalasi, misalnya lonjakan topik “login/OTP” atau “crash” dengan negative share tinggi langsung menjadi tiket prioritas; (3) mengintegrasikan dashboard dengan workflow manajemen isu (misalnya helpdesk/ticketing) agar topik-topik dominan terkonversi menjadi backlog perbaikan yang dapat ditelusuri; (4) melakukan *human-in-the-loop review* untuk topik prioritas agar klasifikasi/topik tetap interpretabel dan mengurangi risiko salah prioritas; serta (5) melakukan pembaruan model secara berkala untuk mengantisipasi drift kosakata dan perubahan pola keluhan setelah pembaruan aplikasi (Panichella et al., 2015; Maalej et al., 2016; Malgaonkar et al., 2022; Dąbrowski et al., 2022). Dengan langkah ini, sistem yang dibangun tidak hanya menjadi prototipe analitik, melainkan mekanisme dukungan keputusan yang berkelanjutan dalam pengelolaan layanan aplikasi.

KESIMPULAN

Penelitian ini menunjukkan bahwa ulasan pengguna di toko aplikasi dapat diolah menjadi masukan yang terstruktur dan siap tindakan melalui integrasi analitik teks (sentimen dan topik) ke dalam sistem informasi berbasis dashboard. Temuan utama memperlihatkan bahwa klasifikasi sentimen berbasis TF-IDF mampu memetakan persepsi pengguna secara andal, dengan model SVM menghasilkan performa terbaik (Accuracy 0,86; Macro-F1 0,84). Pemodelan topik (LDA) berhasil merangkum ulasan menjadi isu dominan dan mengungkap bahwa sentimen negatif paling kuat terkonsentrasi pada masalah yang bersifat “blocker” bagi layanan inti—seperti gangguan login/OTP, crash setelah update, kegagalan pembayaran/transaksi, dan downtime server. Dari sisi adopsi, dashboard yang dibangun memperoleh skor usability tinggi (SUS 82,3), sehingga menunjukkan kesiapan digunakan untuk monitoring rutin, triase, dan pelaporan berkala, serta memperkuat pelajaran bahwa nilai analitik meningkat ketika keluaran NLP dioperasionalkan dalam bentuk sistem informasi yang mudah dipakai pengguna.

Kontribusi ilmiah penelitian ini terletak pada penyediaan pendekatan end-to-end yang menjembatani Teknik Informatika dan Sistem Informasi: mulai dari pengumpulan dan kurasi ulasan, pemodelan sentimen terawasi, ekstraksi isu melalui pemodelan topik, pemetaan topik–sentimen untuk prioritasasi, hingga implementasi dashboard dan evaluasi usability menggunakan instrumen terstandar. Dengan demikian, penelitian ini tidak hanya menyajikan hasil model, tetapi juga menawarkan artefak sistem informasi yang dapat direplikasi, dievaluasi, dan diadaptasi untuk konteks aplikasi layanan informasi lain. Selain itu, penelitian ini mengusulkan kerangka prioritasasi yang praktis—menggabungkan prevalensi topik dan proporsi sentimen negatif—sebagai mekanisme pengambilan keputusan yang lebih objektif untuk backlog pemeliharaan aplikasi.

Meski demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, studi dilakukan pada satu konteks aplikasi sehingga generalisasi ke domain lain perlu diuji lebih lanjut. Kedua, pelabelan sentimen menggunakan rating sebagai proxy berpotensi tidak menangkap ulasan bernuansa campuran (misalnya rating tinggi namun teks berisi keluhan), sehingga akurasi label dapat memengaruhi evaluasi model. Ketiga, pemodelan topik membutuhkan interpretasi manual untuk penamaan topik dan dapat terpengaruh oleh perubahan kosakata (concept drift) setelah pembaruan aplikasi. Penelitian selanjutnya disarankan menggunakan pelabelan manual/ground truth untuk sebagian sampel, menguji model berbasis transformer yang diadaptasi untuk domain software engineering, menambahkan segmentasi berdasarkan versi aplikasi/perangkat untuk mengaitkan isu dengan rilis tertentu, serta melakukan validasi lintas beberapa aplikasi agar manfaat sistem dapat dibuktikan secara lebih komprehensif.

DAFTAR PUSTAKA

- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

- Calefato, F., Lanubile, F., Maiorano, F., & Novielli, N. (2018). Sentiment polarity detection for software development. *Empirical Software Engineering*, 23(3), 1352–1382. <https://doi.org/10.1007/s10664-017-9546-9>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Dąbrowski, J., Letier, E., Perini, A., & Susi, A. (2022). Analysing app reviews for software engineering: A systematic literature review. *Empirical Software Engineering*, 27(2), 43. <https://doi.org/10.1007/s10664-021-10065-7>
- DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, 19(4), 9–30. <https://doi.org/10.1080/07421222.2003.11045748>
- Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software*, 125, 207–219. <https://doi.org/10.1016/j.jss.2016.11.027>
- Islam, M. R., & Zibran, M. F. (2018). SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software*, 145, 125–146. <https://doi.org/10.1016/j.jss.2018.08.030>
- Jongeling, R., Sarkar, P., Datta, S., & Serebrenik, A. (2017). On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering*, 22(5), 2543–2584. <https://doi.org/10.1007/s10664-016-9493-x>
- Lewis, J. R. (2018). The System Usability Scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>
- Lin, B., Cassee, N. W., Serebrenik, A., Bavota, G., Novielli, N., & Lanza, M. (2022). Opinion mining for software development: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 31(3), 38. <https://doi.org/10.1145/3490388>
- Maalej, W., Kurtanović, Z., Nabil, H., & Stanik, C. (2016). On the automatic classification of app reviews. *Requirements Engineering*, 21(3), 311–331. <https://doi.org/10.1007/s00766-016-0251-9>

- Malgaonkar, S., Licorish, S. A., & Savarimuthu, B. T. R. (2022). Prioritizing user concerns in app reviews: A study of requests for new features, enhancements and bug fixes. *Information and Software Technology*, 144, 106798. <https://doi.org/10.1016/j.infsof.2021.106798>
- Martin, W., Sarro, F., Jia, Y., Zhang, Y., & Harman, M. (2017). A survey of app store analysis for software engineering. *IEEE Transactions on Software Engineering*, 43(9), 817–847. <https://doi.org/10.1109/TSE.2016.2630689>
- Novielli, N., Calefato, F., Lanubile, F., & Serebrenik, A. (2021). Assessment of off-the-shelf SE-specific sentiment analysis tools: An extended replication study. *Empirical Software Engineering*, 26, 77. <https://doi.org/10.1007/s10664-021-09960-w>
- Obaidi, M., Nagel, L., Specht, A., & Klünder, J. (2022). Sentiment analysis tools in software engineering: A systematic mapping study. *Information and Software Technology*, 151, 107018. <https://doi.org/10.1016/j.infsof.2022.107018>
- Panichella, S., Di Sorbo, A., Guzman, E., Visaggio, C. A., Canfora, G., & Gall, H. C. (2015). How can I improve my app? Classifying user reviews for software maintenance and evolution. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (pp. 281–290). <https://doi.org/10.1109/ICSME.2015.7332474>
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163–173. <https://doi.org/10.1002/asi.21662>
- Wang, X., Zhang, T., Tan, Y., Shang, W., & Li, Y. (2024). How to effectively mine app reviews concerning software ecosystem? A survey of review characteristics. *Journal of Systems and Software*, 213, 112040. <https://doi.org/10.1016/j.jss.2024.112040>
- Yang, C., Wu, L., Yu, C., & Zhou, Y. (2021). A phrase-level user requests mining approach in mobile application reviews: Concept, framework, and operation. *Information*, 12(5), 177. <https://doi.org/10.3390/info12050177>
- Zhang, T., Xu, B., Thung, F., Haryono, S. A., Lo, D., & Jiang, L. (2020). Sentiment analysis for software engineering: How far can pre-trained transformer models go? In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. <https://doi.org/10.1109/ICSME46990.2020.00017>

