

Analisis Sentimen pada Ulasan Aplikasi FinTech di Indonesia: Studi Komparatif Model Machine Learning dan Deep Learning

Achmad Lutfi Fuadi¹, Agus Heri Yunial², Andrian Hidayat³, Ade Napila⁴, Ahmad Fauzi⁵

¹²³⁴⁵Program Studi Teknik Informatika, Universitas Pamulang, Indonesia

Corresponding Author: Ahmad Fauzi, dosen02621@unpam.ac.id

Published: Januari, 2026

ABSTRAK

Pertumbuhan sektor Teknologi Finansial (FinTech) telah menjadikan umpan balik pelanggan dari platform digital sebagai sumber penting untuk pengambilan keputusan strategis. Namun, volume dan ketidakstrukturan data, khususnya dalam bahasa informal seperti Bahasa Indonesia, menimbulkan tantangan analitis yang signifikan. Penelitian ini bertujuan untuk mengidentifikasi pipeline optimal untuk klasifikasi sentimen pada ulasan pengguna Livin' by Mandiri, super-app perbankan digital Indonesia. Kami melakukan analisis komparatif menggunakan dataset dunia nyata berisi 117.471 ulasan yang tidak seimbang (55% negatif, 31% positif, 14% netral) yang dibersihkan dari Google Play Store. Dua teknik vektorisasi teks, Bag-of-Words (BoW) dan TF-IDF, diuji pada empat classifier machine learning: Random Forest, Logistic Regression, Decision Tree, dan SVM, serta dibandingkan dengan model Deep Learning berbasis Long Short-Term Memory (LSTM). Hasilnya menunjukkan bahwa model LSTM unggul dengan akurasi 98,02% dan weighted F1-score 0,99, sementara model machine learning terbaik, Logistic Regression dengan TF-IDF, menghasilkan weighted F1-score 0,92. Temuan ini menegaskan bahwa meskipun machine learning tradisional efektif, LSTM lebih unggul dalam menangkap konteks dalam data sekuensial yang kompleks dan tidak seimbang. Penelitian ini menawarkan kerangka kerja yang berguna bagi institusi keuangan untuk menerapkan sistem analisis sentimen otomatis yang akurat dan efektif.

Kata Kunci: Analisis Sentimen, FinTech, Deep Learning, LSTM, Machine Learning, TF-IDF

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>

PENDAHULUAN

Digitalisasi sektor keuangan Indonesia telah berkembang pesat, mendorong transformasi layanan perbankan melalui kehadiran aplikasi digital terpadu atau “super-app.” Salah satu contoh menonjol adalah Livin' by Mandiri, yang hingga akhir 2024 telah digunakan oleh lebih dari 29,3 juta pengguna dan memproses 3,9 miliar transaksi dengan pertumbuhan tahunan sebesar 38%. Volume interaksi yang tinggi ini menghasilkan data tekstual dalam jumlah besar, khususnya ulasan pengguna di Google Play Store, yang berpotensi menjadi sumber berharga untuk analisis persepsi pelanggan.

Analisis sentimen, bagian penting dari Natural Language Processing (NLP), bertujuan mengidentifikasi opini dalam data tekstual (Sharma & Goyal, 2023). Dengan ledakan data digital, pemahaman otomatis terhadap sentimen publik menjadi aset strategis di berbagai sektor (Tan et al., 2023). Pendekatan Machine Learning (ML) dan Deep Learning (DL) telah diterapkan luas, dengan keunggulan dan tantangannya, seperti mengatasi sarkasme dan ambiguitas, dimana sarkasme menyamarkan makna sebenarnya di balik ekspresi positif, sedangkan ambiguitas muncul saat kata atau frasa memiliki makna ganda yang bergantung pada konteks, sehingga menyulitkan model dalam menentukan sentimen yang tepat. (Sindhu et al., 2023; Sushma et al., 2024).

Seiring waktu, pendekatan analisis sentimen telah berevolusi dari metode leksikal menuju algoritma pembelajaran mesin (ML) dan pembelajaran dalam (DL) yang lebih kompleks (Du et al., 2021; Sindhu et al., 2023). Model ML klasik seperti Naive Bayes, Decision Tree, dan Support Vector Machine (SVM) terbukti efektif dalam berbagai studi (Ependi et al., 2023; Kavitha et al., 2022). Di sisi lain, model DL seperti Long Short-Term Memory (LSTM) dan BERT lebih unggul dalam

menangkap konteks sekuensial dalam teks panjang (Jazuli et al., 2025; Ma'aly et al., 2024). Namun, efektivitas masing-masing model sangat bergantung pada karakteristik dataset yang digunakan (Alawaji & Aloraini, 2025; Ashbaugh & Zhang, 2024).

Analisis sentimen kini menjadi alat strategis dalam memahami persepsi publik di berbagai sektor. Di media sosial, ia digunakan untuk memantau opini masyarakat secara real-time terhadap kebijakan, isu sosial, dan krisis (Alqahtani et al., 2023; Nip & Berthelie, 2024). Dalam politik, teknik ini membantu membaca arah dukungan pemilih dan respons terhadap isu pemilu (Ma'aly et al., 2024). sementara di sektor komersial, digunakan untuk mengevaluasi kepuasan pelanggan dan tren pasar dari ulasan e-commerce (Ghatora et al., 2024; Yang, 2023).

Penerapannya juga meluas ke sektor pariwisata, transportasi, smart city, dan kesehatan digital mulai dari menilai pengalaman wisatawan hingga memantau reaksi pengguna terhadap layanan publik dan aplikasi (Ependi et al., 2023; Kim & Chung, 2024; Refianti et al., 2024; Torres & de Picado-Santos, 2025; Zakarija et al., 2024). Hal ini menunjukkan bahwa analisis sentimen tak hanya bersifat evaluatif, tetapi juga berfungsi sebagai sumber wawasan penting untuk inovasi, kebijakan, dan peningkatan kualitas layanan lintas domain.

Namun, terlepas dari luasnya penerapan tersebut, masih terdapat beberapa *gap* krusial yang membatasi efektivitas analisis sentimen, khususnya dalam konteks teks berbahasa lokal yang informal. Tantangan utama mencakup proses pra-pemrosesan data untuk bahasa yang multibahasa dan tidak baku, seperti yang sering ditemukan dalam ulasan pengguna di platform digital, yang dapat menurunkan akurasi model (Garg & Sharma, 2022). Selain itu, terdapat pula hambatan dalam akses data berkualitas serta isu etika dalam pengumpulan dan pemanfaatan data pengguna (Nip & Berthelie, 2024).

Di Indonesia, studi analisis sentimen mulai mengadopsi model-model canggih seperti IndoBERT dan LSTM berbasis leksikon untuk aplikasi tertentu (Jazuli et al., 2025; Kim & Chung, 2024). Namun, hingga saat ini belum terdapat penelitian yang secara komprehensif membandingkan performa model machine learning (ML) klasik dan deep learning (DL) khususnya arsitektur LSTM dalam mengolah ulasan pengguna berskala besar (>100.000 data) pada super-app finansial seperti Livin'. Untuk mengisi kesenjangan tersebut, penelitian ini melakukan analisis komparatif guna mengidentifikasi pipeline klasifikasi sentimen yang optimal di industri FinTech Indonesia. Fokus utama penelitian ini adalah membandingkan performa model ML klasik (menggunakan metode vektorisasi TF-IDF dan Bag of Words) dengan model DL berbasis LSTM yang memanfaatkan word embedding. Selain itu, penelitian ini menguji efektivitas tahapan pra-pemrosesan berbasis lokal, termasuk normalisasi dengan kamus slang Bahasa Indonesia, stemming menggunakan algoritma Sastrawi, serta analisis frekuensi kata dengan FreqDist. Dengan memanfaatkan dataset besar dan tidak seimbang (>100.000 ulasan berbahasa Indonesia informal), studi ini diharapkan dapat menghasilkan pendekatan optimal dalam analisis sentimen serta menawarkan kerangka kerja strategis bagi institusi keuangan dalam merancang sistem umpan balik pelanggan yang akurat dan efisien, sekaligus berkontribusi pada pengembangan sistem NLP di sektor FinTech Indonesia.

TINJAUAN PUSTAKA

Analisis sentimen, atau opinion mining, adalah bidang studi interdisipliner yang memanfaatkan pemrosesan bahasa alami, machine learning, dan linguistik komputasi untuk mengidentifikasi, mengekstrak, dan mengkuantifikasi informasi subjektif dalam teks (Refianti et al., 2024). Analisis sentimen telah menjadi alat yang sangat diperlukan bagi bisnis dan organisasi untuk memahami sentimen pelanggan, memantau reputasi merek, dan membuat keputusan berdasarkan data (Tan et al., 2023). Dalam era big data, pertumbuhan eksponensial data tekstual telah memicu adopsi luas analisis sentimen, memungkinkan identifikasi otomatis polaritas sentimen dari berbagai sumber data dalam skala besar, dan memfasilitasi klasifikasi serta pemahaman opini publik secara efisien (Kim & Chung, 2024; Sindhu et al., 2023).

Analisis Sentimen dalam Industri FinTech

Dalam industri FinTech yang sangat kompetitif, pengalaman pengguna telah menjadi faktor kunci keberhasilan bisnis, dengan analisis data sebagai inti untuk beralih dari layanan pelanggan

reaktif ke proaktif. Analisis sentimen, seperti yang dijelaskan oleh (Sindhu et al., 2023; Sushma et al., 2024), kini menjadi instrumen strategis yang memungkinkan perusahaan untuk memantau reputasi, mengidentifikasi masalah, dan menilai peluncuran fitur baru secara real-time (Sharma & Goyal, 2023). Prinsip dasar machine learning, "garbage in, garbage out," menekankan pentingnya kualitas data, sehingga tahap pra-pemrosesan dan pelabelan data sangat menentukan keberhasilan analisis sentimen, khususnya untuk teks Bahasa Indonesia informal. (Garg & Sharma, 2022) menekankan bahwa pra-pemrosesan yang efektif meningkatkan akurasi klasifikasi secara signifikan.

Pra-pemrosesan Teks Bahasa Indonesia

Teks ulasan pengguna Indonesia sering mengandung "noise" seperti singkatan non-standar, slang, dan inkonsistensi ejaan. Untuk mengatasi ini, dua langkah penting adalah normalisasi dan stemming. Normalisasi menggunakan kamus slang untuk mengubah kata informal menjadi bentuk baku, sementara stemming, melalui pustaka Sastrawi, mengubah kata berimbuhan menjadi kata dasar untuk mengurangi redundansi fitur. Pelabelan sentimen, meskipun pelabelan manual dianggap akurat, sangat mahal dan tidak praktis untuk dataset besar (>100.000 ulasan). Sebagai alternatif, pendekatan berbasis leksikon (Ependi et al., 2023) digunakan untuk memberikan sentimen otomatis berdasarkan kamus kata positif dan negatif. Meskipun terbatas dalam menangani nuansa domain spesifik, penelitian seperti (Du et al., 2021) menunjukkan bahwa pendekatan ini efektif untuk analisis data besar, dengan keseimbangan yang baik antara kecepatan dan kelayakan.

Tantangan Keseimbangan Data dan Vektorisasi Fitur

Dengan menganalisis ulasan dari toko aplikasi, perusahaan dapat memperoleh wawasan penting untuk pengembangan produk dan strategi pemasaran (Ghatora et al., 2024), terutama di sektor perbankan digital (Alawaji & Aloraini, 2025). Namun, data ini seringkali memiliki distribusi kelas yang tidak seimbang, dengan ulasan negatif atau positif lebih dominan (Kim & Chung, 2024). Beberapa penelitian mengatasi masalah ini menggunakan teknik resampling atau oversampling untuk menghindari bias terhadap kelas mayoritas (Alqahtani et al., 2023; Zakariya et al., 2024), menunjukkan pentingnya penanganan data tidak seimbang untuk analisis valid. Dalam konteks ekonomi digital Indonesia yang berkembang pesat, kemampuan untuk mengelola dan bertindak cepat berdasarkan sentimen nasabah menjadi keunggulan kompetitif. Setelah data dibersihkan dan diberi label, vektorisasi dilakukan dengan model Bag-of-Words (BoW) dan Term Frequency-Inverse Document Frequency (TF-IDF), dengan TF-IDF lebih sering digunakan karena kemampuannya memberikan bobot lebih pada kata-kata informatif (Alawaji & Aloraini, 2025; Kavitha et al., 2022).

Model Machine Learning untuk Analisis Sentimen

Meskipun teknik seperti Bag-of-Words dan TF-IDF penting (Garg & Sharma, 2022), keduanya memiliki kelemahan utama: mereka mengabaikan urutan kata, kehilangan konteks gramatikal dan semantik. Hal ini memotivasi peralihan ke model yang lebih canggih. Machine learning menawarkan berbagai algoritma yang dapat dilatih untuk mengklasifikasikan sentimen dari data tekstual. Berbagai classifier Machine Learning (ML) seperti Random Forest, SVM, dan Logistic Regression telah menunjukkan kinerja solid dalam analisis sentimen. (Kavitha et al., 2022) melaporkan akurasi 95% dengan Random Forest, sementara (Sharma & Goyal, 2023) mencatat akurasi 94,05% dengan SVM. Namun, untuk memastikan hasil yang andal, validasi silang (cross-validation) diperlukan, memberikan estimasi kinerja yang lebih stabil dan dapat digeneralisasi. Dengan validasi yang kuat, model ML menetapkan baseline kinerja, di mana metode baru harus membuktikan keunggulannya secara signifikan.

Model Deep Learning dan NLP Modern

Kebangkitan Deep Learning (DL) menandai pergeseran paradigma dalam NLP, dengan Long Short-Term Memory (LSTM) yang dirancang untuk mengatasi kelemahan model klasik melalui pemrosesan data sekuensial dan kemampuan mengingat konteks jangka panjang (Sindhu et al., 2023). Kemampuan ini sangat penting untuk analisis sentimen yang akurat. Bukti empiris dalam Bahasa Indonesia menunjukkan efektivitas LSTM dan model canggih lainnya. (Ma'aly et al., 2024)

mencapai akurasi 91% dengan Bi-LSTM pada sentimen multi-label, (Jazuli et al., 2025) menggunakan IndoBERT dengan akurasi 97,9%, dan (Refianti et al., 2024) memperoleh akurasi 85,3% dengan LSTM pada ulasan aplikasi kesehatan. Meskipun model ML klasik seperti yang dilaporkan oleh (Kavitha et al., 2022; Sharma & Goyal, 2023) memberikan baseline yang solid, arsitektur DL seperti LSTM dan BERT menawarkan pemahaman kontekstual lebih mendalam (Jazuli et al., 2025; Ma'aly et al., 2024). Namun, temuan yang berlawanan, seperti yang ditemukan oleh (Alawaji & Aloraini, 2025; Ashbaugh & Zhang, 2024), menunjukkan bahwa model ML klasik dapat mengungguli DL, menyoroti pentingnya konteks dalam memilih model yang optimal.

Gap Penelitian dan Kontribusi Studi

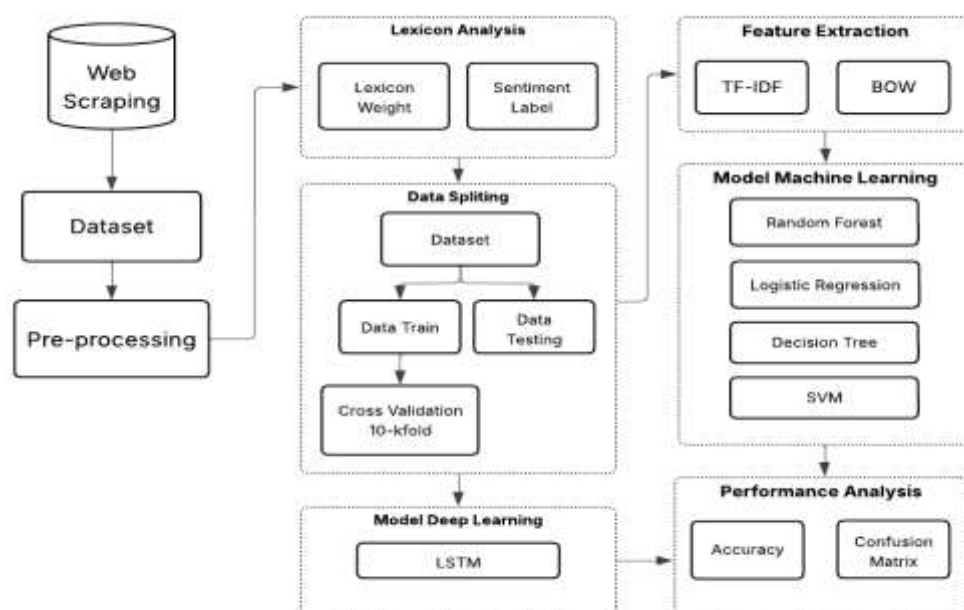
Meskipun terdapat studi di sektor perbankan untuk bahasa Arab (Alawaji & Aloraini, 2025) dan penerapan DL di domain lain di Indonesia (Refianti et al., 2024), belum ada studi yang secara sistematis dan komprehensif membandingkan model ML klasik dan arsitektur DL (LSTM) pada dataset super-app finansial Indonesia yang besar, tidak seimbang, dan divalidasi dengan cross-validation. Menunjukkan bahwa belum ada studi yang membandingkan performa ML dan DL secara komprehensif terhadap ulasan FinTech berskala besar (>100.000), serta pentingnya pendekatan lokal dalam preprocessing dan model evaluasi.

METODOLOGI PENELITIAN

Studi ini menerapkan metodologi komprehensif yang mencakup pengumpulan data, pra-pemrosesan, vektorisasi, pelatihan model, dan evaluasi. Data diperoleh dari ulasan publik pengguna aplikasi perbankan digital Indonesia di Google Play Store, yang mencerminkan beragam pengalaman terkait kegunaan, fitur, kinerja, dan layanan pelanggan.

Kerangka Penelitian

Penelitian ini dilakukan melalui tahapan terstruktur yang meliputi: pengumpulan data ulasan aplikasi *Livin' by Mandiri* dari Google Play Store, pra-pemrosesan teks (termasuk normalisasi dan *stemming* Sastrawi), pelabelan sentimen otomatis berbasis leksikon, vektorisasi menggunakan BoW dan TF-IDF, pelatihan serta validasi lima model (empat ML dan satu DL) dengan *10-fold cross-validation*, dan evaluasi performa menggunakan metrik standar untuk menentukan model terbaik. Diagram alur penelitian ditampilkan pada Gambar 1.



Gambar 1. Desain Sistem Penelitian

Pengumpulan Data

Penelitian ini menggunakan data ulasan pengguna aplikasi Livin' by Mandiri yang dikumpulkan dari platform Google Play Store melalui bantuan pustaka google-play-scraper pada rentang waktu 30 September 2021 hingga 12 Juli 2024. Dari total 139.500 ulasan yang diperoleh, dilakukan proses data cleaning untuk menghilangkan duplikasi, komentar kosong, dan entri yang tidak relevan, sehingga menghasilkan 117.471 ulasan akhir yang siap dianalisis. Ulasan yang digunakan ditulis dalam bahasa Indonesia informal dan mencerminkan pengalaman pengguna yang beragam. Pemilihan aplikasi Livin' didasarkan pada popularitasnya sebagai salah satu super-app finansial terbesar di Indonesia serta volume ulasan yang tinggi.

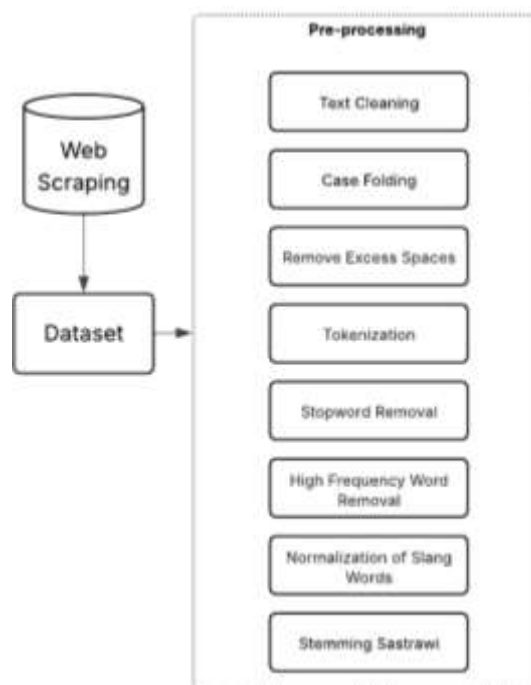
Data bersifat imbalanced, dengan distribusi sentimen tidak merata, yang mencerminkan kondisi sebenarnya dari umpan balik pengguna di aplikasi keuangan digital. Beberapa contoh ulasan ditampilkan pada Tabel 1 untuk memberikan gambaran mengenai jenis komentar yang dianalisis dalam studi ini.

Tabel 1. Komentar pengguna Aplikasi Play Store

No	Content
1	Apa ini... Kita sdh instal sampe 8 kali, tpi nggak bisa masuk aplikasi juga... Sdh di suruh force close aplikasinya, restart hp, hapus cache, pakai wifi, pakai paket data, hapus aplikasi lain.. Tapi tetap sama saja tidak bisa masuk... Mohon solusinya... Aplikasi ini sangat membantu, di saat urgent skli tanpa harus ke atm krna jarak atm terlalu jauh.
2	Jelek banget apk tiba2 mental ga bisa masuk sama sekali ,tlp cs email dll ga ada penyelesaian bilang nya layanan normal ,sudah di force close ,clear cache ,memori hp masih kosong banyak ,instal ulang sudah, masuk tidak memakai jaringan wifi, semua sudah di lakukan tetap mental,malah bilang layanan normal ,tp review kebanyakan ternyata sama kendala nya ,minta solusinya aja gimana karena apk nya sejam sebelum masih bisa di gunakan transaksi ketika mau di gunakan lg tiba2 close sendiri mental mulu
3	Ini aplikasi buat masuk aja gak bisa. Padahal wifi stabil, kondisi hp juga stabil tidak lag entah kenapa makin lama makin lama load untuk masuk, beruntung kalo bisa kadang malah gabisa sama sekali
4	Saya sebagai nasabah bank mandiri, apk nya baguss sihh mau TF, top up apapun juga lancar, login juga lancar2 aja. Cuma gak ada fitur livin paylaternya aja sihh, kalo KPRmh udah muncul. Selebihnya baguss sih cukup membantu dalam transaksi pembayaran VA
5	Gua tf aja suruh periksa pengaturan waktu otomatis, padahal sudah otomatis gk ada masalah, dasar aplikasinya aja yg minus, banyak error, aplikasi sampah, kebanyakan gangguan, sulit di gunakan pas lagi butuh butuhnya.

Pra-pemrosesan Data

Untuk memastikan kualitas data input, serangkaian 8 langkah pra-pemrosesan yang komprehensif diterapkan pada setiap ulasan. Rangkaian langkah ini sangat penting, karena literatur secara konsisten menunjukkan bahwa pembersihan teks yang cermat secara signifikan meningkatkan akurasi klasifikasi akhir dan mengurangi dimensi data yang tidak perlu (Garg & Sharma, 2022). Langkah-langkah tersebut adalah seperti pada gambar 2 berikut:



Gambar 2. Data Pra-pemrosesan

Pada tahap akhir adalah *stemming* menggunakan Sastrawi guna menyederhanakan kata ke bentuk dasar, yang terbukti meningkatkan efektivitas klasifikasi. Tabel 2 perbandingan teks yang sudah dilakukan stemming:

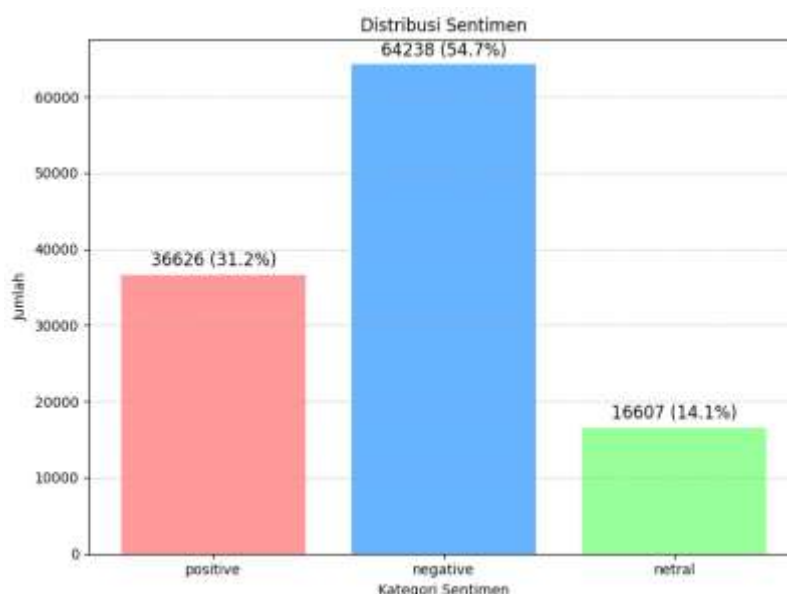
Tabel 2. Perbandingan Hasil Stemming Teks Komentar Menggunakan Sastrawi

No	Content	Stemmed_Content
1	sudah instal sampai kali tapi tidak masuk aplikasi sudah suruh force close aplikasinya restart hp hapus cache pakai wifi pakai paket data hapus aplikasi masuk mohon solusinya aplikasi membantu urgent skli atm karena jarak atm jelek sekali apk mental tidak masuk telepon cs email dan lain tidak penyelesaian bilang nya layanan	sudah instal sampai kali tapi tidak masuk aplikasi sudah suruh force close aplikasi restart hp hapus cache pakai wifi pakai paket data hapus aplikasi masuk mohon solusi aplikasi bantu urgent skli atm karena jarak atm jelek sekali apk mental tidak masuk telepon cs email dan lain tidak selesai bilang nya layan
2	normal force close clear cache memori hp kosong instal ulang masuk memakai jaringan wifi lakukan mentalmalah bilang layanan normal tapi review kebanyakan kendala nya solusinya saja gimana apk nya sejam transaksi lagi close mental terus	normal force close clear cache memori hp kosong instal ulang masuk pakai jaring wifi laku mentalmalah bilang layan normal tapi review banyak kendala nya solusi saja gimana apk nya jam transaksi lagi close mental terus
3	aplikasi masuk saja tidak wifi stabil kondisi hp stabil lag load masuk beruntung kalau kadang tidak bisa nasabah bank mandiri apk nya baguss sihh transfer	aplikasi masuk saja tidak wifi stabil kondisi hp stabil lag load masuk untung kalau kadang tidak bisa nasabah bank mandiri apk nya baguss sihh transfer
4	top up apapun lancar login lancar saja tidak fitur livin paylaternya saja sih kalau kprmh sudah muncul selebihnya baguss sih membantu transaksi pembayaran va	top up apa lancar login lancar saja tidak fitur livin paylaternya saja sih kalau kprmh sudah muncul lebih baguss sih bantu transaksi bayar va
5	saya transfer saja suruh periksa pengaturan otomatis otomatis tidak dasar aplikasinya saja yang minus error aplikasi sampah kebanyakan gangguan sulit pas butuh butuhnya	saya transfer saja suruh periksa atur otomatis otomatis tidak dasar aplikasi saja yang minus error aplikasi sampah banyak ganggu sulit pas butuh butuh

Pelabelan Sentimen

Pelabelan sentimen dilakukan secara otomatis menggunakan pendekatan leksikon karena efisien untuk data besar dan telah divalidasi dalam studi terkait aplikasi seluler dan smart city (Ependi et al., 2023; Refianti et al., 2024). Selain itu, pengembangan leksikon untuk domain spesifik merupakan area penelitian aktif yang diakui untuk meningkatkan akurasi (Du et al., 2021). Dua kamus sentimen, *lexicon_positive.csv* dan *lexicon_negative.csv* (Angel Metanosa Afinda, n.d.),

digunakan untuk memberikan skor pada setiap ulasan, dengan aturan: skor > 0 = positif, < 0 = negatif, dan $= 0$ = netral. Hasilnya dapat dilihat pada gambar 3 berikut:



Gambar 3. Distribusi Sentimen dalam Himpunan Data Setelah Pelabelan

Distribusi ini menggambarkan ketimpangan sentimen yang lazim ditemukan dalam evaluasi pengguna, di mana 64.238 ulasan negatif (54,7%), 36.626 positif (31,2%), dan 16.607 netral (14,1%), sentimen negatif mendominasi, yang merupakan hal yang umum terjadi dalam komentar di toko aplikasi. Tabel 3 menunjukkan distribusi lima baris awal sentimen dalam dataset setelah pelabelan berbasis leksikon.

Tabel 3. Distribusi Sentimen Dalam Himpunan Data Setelah Pelabelan Berbasis Leksikon

No	Stemmed_Content	Polarity_Score	Polarity
1	sudah instal sampai kali tapi tidak masuk aplikasi sudah suruh force close aplikasi restart hp hapus cache pakai wifi pakai paket data hapus aplikasi masuk mohon solusi aplikasi bantu urgent skli atm karena jarak atm	-37	negative
2	jelek sekali apk mental tidak masuk telepon cs email dan lain tidak selesai bilang nya layan normal force close clear cache memori hp kosong instal ulang masuk pakai jaring wifi laku mentalmalah bilang layan normal tapi review banyak kendala nya solusi saja gimana apk nya jam transaksi lagi close mental terus	-30	negative
3	aplikasi masuk saja tidak wifi stabil kondisi hp stabil lag load masuk untung kalau kadang tidak bisa	-13	negative
4	nasabah bank mandiri apk nya baguss sihh transfer top up apa lancar login lancar saja tidak fitur livin paylaternya saja sihh kalau kprmh sudah muncul lebih baguss sihh bantu transaksi bayar va	-6	negative
5	saya transfer saja suruh periksa atur otomatis otomatis tidak dasar aplikasi saja yang minus error aplikasi sampah banyak ganggu sulit pas butuh butuh	-51	negative

Vektorisasi Teks

Dalam penelitian ini, dua metode vektorisasi digunakan untuk merepresentasikan teks: Bag-of-Words (BoW) dan Term Frequency-Inverse Document Frequency (TF-IDF). BoW merepresentasikan setiap dokumen sebagai vektor frekuensi kata dalam kosakata, menawarkan pendekatan sederhana namun efektif sebagai baseline untuk klasifikasi teks. Sementara itu, TF-IDF memberikan bobot lebih tinggi pada kata-kata penting dengan menyesuaikan frekuensi lokal dan global dalam korpus, sehingga meningkatkan akurasi model dalam tugas analisis sentimen (Alawaji & Aloraini, 2025; Kavitha et al., 2022). Keduanya diimplementasikan menggunakan *CountVectorizer* dan *TfidfVectorizer* dari scikit-learn, dengan parameter yang identik ($\text{max_features}=200$, $\text{min_df}=17$, $\text{max_idf}=0.8$) untuk memastikan perbandingan yang adil.

Pemodelan Machine Learning

Penelitian ini menggunakan empat algoritma machine learning populer: Random Forest, Logistic Regression, Decision Tree, dan Support Vector Machine (SVM). Random Forest menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting, sementara Logistic Regression memanfaatkan fungsi sigmoid untuk klasifikasi biner. Decision Tree membagi ruang fitur berdasarkan atribut paling informatif dalam struktur pohon, dan SVM mencari hyperplane optimal untuk memisahkan kelas dengan margin maksimum. Seluruh model diimplementasikan menggunakan parameter default dari scikit-learn, dengan rasio data pelatihan dan pengujian 75:25, serta validasi melalui 10-fold cross-validation. Pemilihan algoritma ini didasarkan pada performa kuat dan rekam jejaknya dalam studi literatur (Alawaji & Aloraini, 2025; Kavitha et al., 2022).

Pemodelan Deep Learning (DL)

Penelitian ini mengembangkan model deep learning berbasis LSTM untuk klasifikasi ulasan aplikasi berbahasa Indonesia, dengan keunggulan dalam memahami konteks sekuensial (Sindhu et al., 2023; Tan et al., 2023). Proses pra-pemrosesan mencakup tokenisasi 5.000 kata teratas, padding hingga 200 token, dan one-hot encoding untuk label sentimen. Model dilatih menggunakan *categorical_crossentropy* dan *adam optimizer*, dengan *batch size* 64 selama 10 *epochs*, serta divalidasi menggunakan 10-fold cross-validation. Arsitektur detail model disajikan pada Tabel 4.

Tabel 4. Detail Arsitektur Model LSTM Yang Digunakan Untuk Klasifikasi Sentimen

Layer	Function	Key Parameters
Embedding Layer	Mengubah indeks kata menjadi representasi vektor padat yang menangkap hubungan semantik.	Vocabulary size: 5,000 Embedding dimension: 128
SpatialDropout1D	Mengatur model dengan menjatuhkan 50% unit penyematan secara acak untuk mencegah pemasangan yang berlebihan.	Dropout rate: 0.5
LSTM Layer	Memproses data berurutan dengan unit memori yang menangkap dependensi jangka panjang antar kata.	Units: 100 Dropout: 0.5 Recurrent dropout: 0.5
Dense Output Layer	Menghasilkan probabilitas kelas sentimen menggunakan aktivasi softmax lebih dari 3 neuron keluaran.	Units: 3 (positive, neutral, negative) Activation: softmax

Metrik Evaluasi

Efektivitas model Machine Learning dan LSTM dalam klasifikasi sentimen ulasan berbahasa Indonesia dievaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score, yang memberikan gambaran menyeluruh atas performa model dalam mengklasifikasikan data secara akurat, baik per kategori sentimen maupun secara keseluruhan. Secara matematis, akurasi didefinisikan sebagai:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Selain metrik numerik *confusion matrix* memberikan rincian yang lebih mendalam tentang prediksi benar dan salah untuk setiap kelas. Untuk klasifikasi biner, matriks ini direpresentasikan sebagai:

$$Confusion\ Matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (5)$$

HASIL DAN ANALISIS

Bagian ini menyajikan hasil kuantitatif dari seluruh eksperimen yang dilakukan, dimulai dari hasil pelabelan sentimen hingga evaluasi performa berbagai model klasifikasi. Berdasarkan proses pelabelan otomatis menggunakan pendekatan leksikon, diperoleh distribusi sentimen dari 117.471 ulasan pengguna aplikasi Livin' by Mandiri yang terdiri atas 36.626 ulasan positif (31,2%), 64.238 ulasan negatif (54,7%), dan 16.607 ulasan netral (14,1%). Distribusi ini memperlihatkan ketidakseimbangan kelas yang cukup signifikan, dengan dominasi ulasan negatif yang mencerminkan kecenderungan pengguna untuk lebih aktif memberikan respons saat mengalami pengalaman buruk. Kondisi ini menjadi pertimbangan penting dalam tahapan pelatihan dan evaluasi model klasifikasi, karena dapat memengaruhi akurasi prediksi serta ketahanan model terhadap bias kelas mayoritas.

Hasil Eksperimen Machine Learning

Empat algoritma *machine learning* dievaluasi menggunakan dua skema vektorisasi TF-IDF dan Bag-of-Words (BoW). Kinerja masing-masing model dianalisis secara individual di bawah ini.

Kinerja Model Random Forest

Random Forest menunjukkan kinerja yang baik dengan akurasi pengujian sebesar 90.23% (TF-IDF) dan 89.86% (BoW). Namun, terdapat indikasi *overfitting* karena selisih yang cukup besar dengan akurasi pelatihannya (96.4%). Berikut Tabel 5 Perbandingan hasil Kinerja Random Forest:

Tabel 5. Perbandingan Hasil Akurasi Random Forest

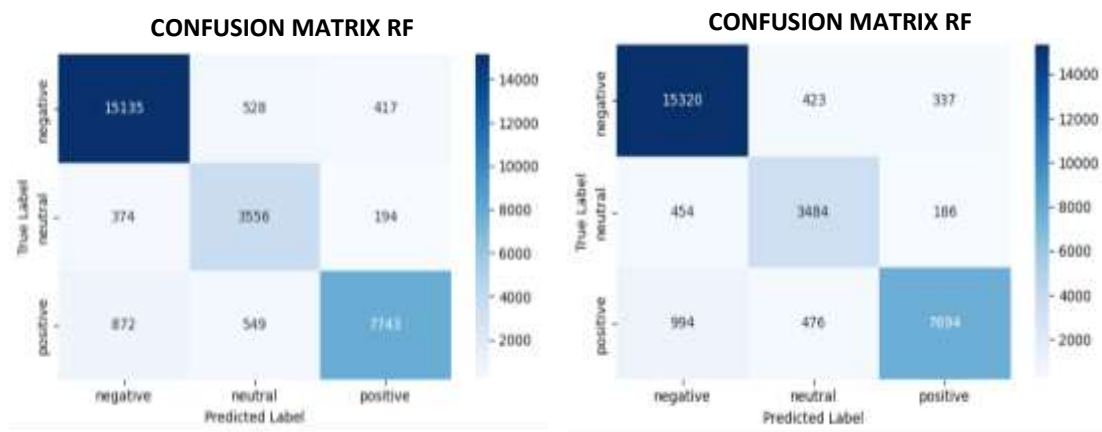
Model	Vektorisasi	Akurasi Pelatihan	Akurasi Pengujian
Random Forest	TF-IDF	96.44%	90.23%
	BoW	96.46%	89.86%

Analisis lebih dalam pada *Classification Report* untuk varian terbaiknya (TF-IDF) menunjukkan pola yang menarik. Model ini sangat andal dalam mengidentifikasi ulasan negatif, dengan *recall* mencapai 0.95. Namun, kinerjanya menurun saat mengklasifikasikan ulasan netral (*recall* 0.84) dan positif (*recall* 0.84). Ini menandakan bahwa Random Forest dalam konfigurasi ini memiliki kecenderungan untuk lebih "mengenal" pola kata-kata negatif, tetapi kurang sensitif terhadap nuansa pada kelas lainnya. Berikut tabel 6 perbandingan hasil confusion matrix random forest:

Tabel 6. Perbandingan Hasil Confusion Matrix Random Forest

Method	Accuracy		Precision		Recall		F1-Score	
	Train	Test	Macro	Weighted	Macro	Weighted	Macro	Weighted
Random Forest + TF-IDF	96.44%	90.2%	88.0%	90.00%	88.0%	90.00%	88.0%	90.00%
Random Forest + BoW	96.46%	89.9%	87.3%	90.00%	88.0%	90.00%	87.3%	90.00%

Gambar 4 menunjukkan *confusion matrix* model Random Forest dengan pendekatan TF-IDF dan BoW, keduanya mencapai akurasi 90% dengan dominasi prediksi tepat pada kelas *negative*. Meskipun demikian, performa pada kelas *neutral* dan *positive* masih menghadapi tantangan, terutama pada *precision* untuk *neutral* dan *recall* untuk *positive*. Pendekatan TF-IDF menunjukkan keseimbangan yang lebih baik antar kelas, sementara BoW cenderung menghasilkan *recall* lebih tinggi pada *neutral* namun kurang presisi. Perbedaan ini mencerminkan kesulitan model dalam membedakan nuansa linguistik antara sentimen netral dan negatif ringan dalam ulasan pengguna.



Gambar 4. Confusion Matrix Random Forest

Kinerja Model Logistic Regression

Logistic Regression secara konsisten menunjukkan kinerja paling stabil. Model ini mencapai akurasi pengujian tertinggi sebesar 91.56% (TF-IDF) dan 91.27% (BoW), dengan tingkat *overfitting* yang sangat rendah. Berikut tabel 7 hasil akurasi model Logistic Regression:

Tabel 7. Perbandingan Hasil Akurasi Logistic Regression

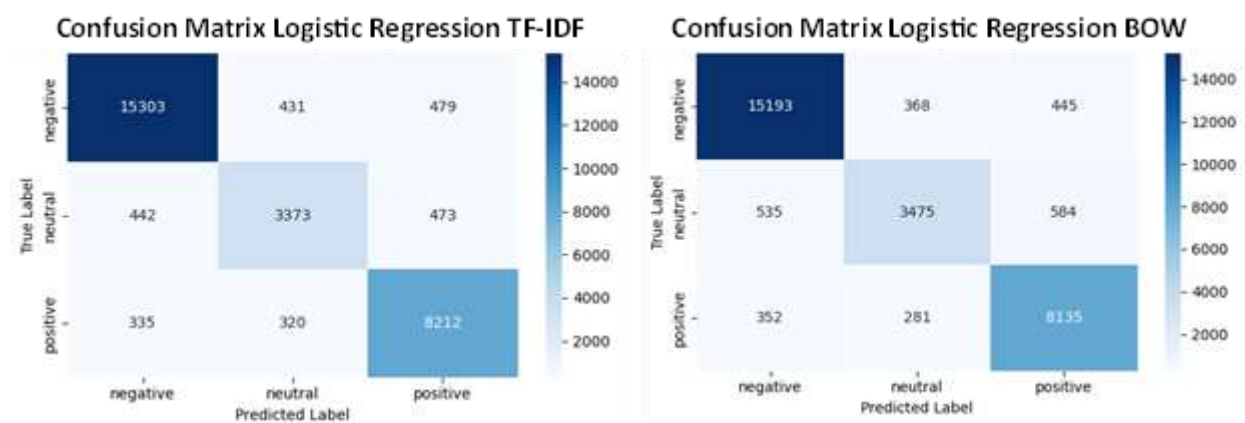
Model	Vektorisasi	Akurasi Pelatihan	Akurasi Pengujian
Logistic Regression	TF-IDF	91.66%	91.56%
	BoW	91.54%	91.27%

Classification Report untuk varian TF-IDF menunjukkan performa yang sangat seimbang dan kuat untuk kelas mayoritas. Baik untuk ulasan negatif (*recall* 0.94) maupun positif (*recall* 0.93), model ini menunjukkan kemampuan klasifikasi yang sangat baik. Namun, kelemahan utamanya yang paling jelas juga terungkap di sini: model ini kesulitan mengidentifikasi ulasan netral, dengan nilai *recall* yang turun signifikan ke angka 0.79. Ini adalah pola yang akan kita lihat pada model ML lainnya. Hasil compusion matrix dapat dilihat pada tabel 8 berikut:

Tabel 8. Perbandingan Hasil Confusion Matrix Logistic Regression

Method	Accuracy		Precision		Recall		F1-Score	
	Train	Test	Macro	Weighted	Macro	Weighted	Macro	Weighted
Logistic Regression + TF-IDF	96.44%	90.2%	88.0%	90.00%	88.0%	90.00%	88.0%	90.00%
Logistic Regression + BoW	96.46%	89.9%	87.3%	90.00%	88.0%	90.00%	87.3%	90.00%

Gambar 5 hasil *confusion matrix* untuk model Logistic Regression dengan TF-IDF dan BoW menunjukkan akurasi tinggi masing-masing 92% dan 91%. Kedua pendekatan unggul dalam mengklasifikasikan sentimen *negative* dan *positive*, dengan nilai *precision* dan *recall* yang konsisten di atas 0.90. Namun, sentimen *neutral* masih menjadi tantangan, terutama pada pendekatan BoW yang menghasilkan *recall* lebih rendah (0.76). Pendekatan TF-IDF menunjukkan keseimbangan kinerja antar kelas yang lebih stabil dibandingkan BoW. Hal ini mengindikasikan bahwa representasi fitur TF-IDF mampu menangkap nuansa teks dengan lebih baik dalam konteks klasifikasi sentimen yang melibatkan ekspresi netral.



Gambar 5. Confusion Matrix Logistic Regression

Kinerja Model Decision Tree

Model Decision Tree tunggal, seperti yang diperkirakan, menunjukkan kinerja terendah dengan akurasi pengujian 87.39% (TF-IDF) dan 87.24% (BoW), serta menunjukkan *overfitting* yang parah. Berikut tabel 9 hasil akurasi model Decision Tree:

Tabel 9. Perbandingan Hasil Akurasi Decision Tree

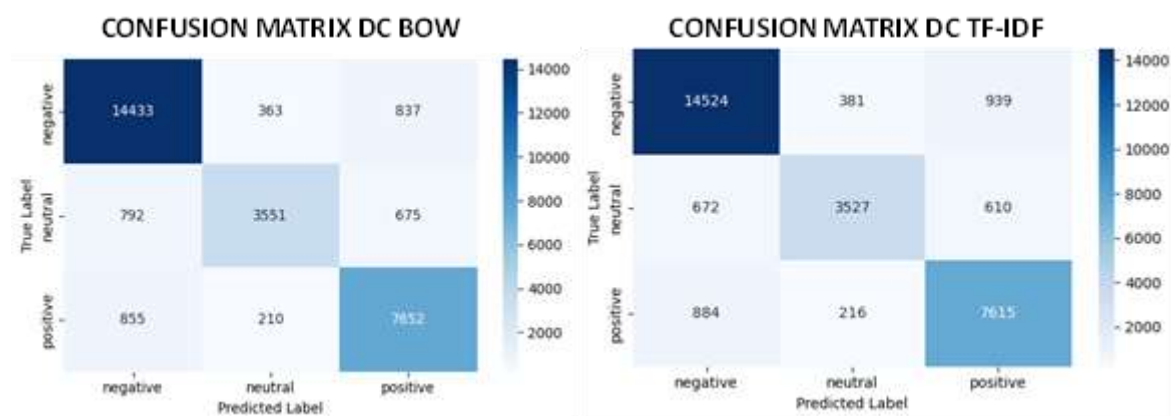
Model	Vektorisasi	Akurasi Pelatihan	Akurasi Pengujian
Decision Tree	TF-IDF	96.44%	87.39%
	BoW	96.46%	87.24%

Pola kesalahannya adalah yang paling ekstrem. Analisis *confusion matrix*-nya menunjukkan bahwa model ini tidak hanya kesulitan pada kelas netral (*recall* 0.73), tetapi juga pada kelas positif (*recall* 0.87), meskipun masih relatif baik pada kelas negatif (*recall* 0.92). Ini mengonfirmasi bahwa tanpa teknik *ensemble* (seperti pada Random Forest), satu Decision Tree tidak cukup kuat untuk menangani kompleksitas data ini. Berikut tabel 10 hasil confusion matrix model Decision Tree:

Tabel 10. Perbandingan Hasil Confusion Matrix Decision Tree

Method	Accuracy		Precision		Recall		F1-Score	
	Train	Test	Macro	Weighted	Macro	Weighted	Macro	Weighted
Decision Tree + TF-IDF	96.44%	87.4%	86.0%	89.00%	87.0%	89.00%	86.5%	89.00%
Decision Tree + BoW	96.46%	87.2%	85.5%	88.00%	86.0%	88.00%	85.7%	88.00%

Gambar 6 hasil model Decision Tree yang dikombinasikan dengan TF-IDF dan BoW menunjukkan performa klasifikasi yang seimbang dengan akurasi keseluruhan sebesar 87% pada kedua pendekatan. Sentimen *negative* dan *positive* diklasifikasikan dengan baik, terlihat dari nilai *precision* dan *recall* di atas 0.90 untuk *negative* dan mendekati 0.90 untuk *positive*. Namun, sentimen *neutral* masih menghadapi tantangan, khususnya pada aspek *recall* yang hanya mencapai 0.73 dengan TF-IDF dan 0.71 dengan BoW. Meskipun begitu, nilai *macro* dan *weighted average* f1-score tetap stabil di angka 0.85 dan 0.87, mengindikasikan kinerja model yang konsisten di seluruh kelas dengan sedikit keunggulan pada fitur representasi TF-IDF.



Gambar 6. Confusion Matrix Decision Tree

Kinerja Model Support Vector Machine (SVM)

SVM dengan kernel linear menunjukkan performa yang sangat kompetitif, setara dengan Logistic Regression, dengan akurasi pengujian 92% dan *weighted F1-score* 0.91 untuk varian TF-IDF. Berikut tabel 11 hasil akurasi model SVM:

Tabel 11. Perbandingan Hasil Akurasi SVM

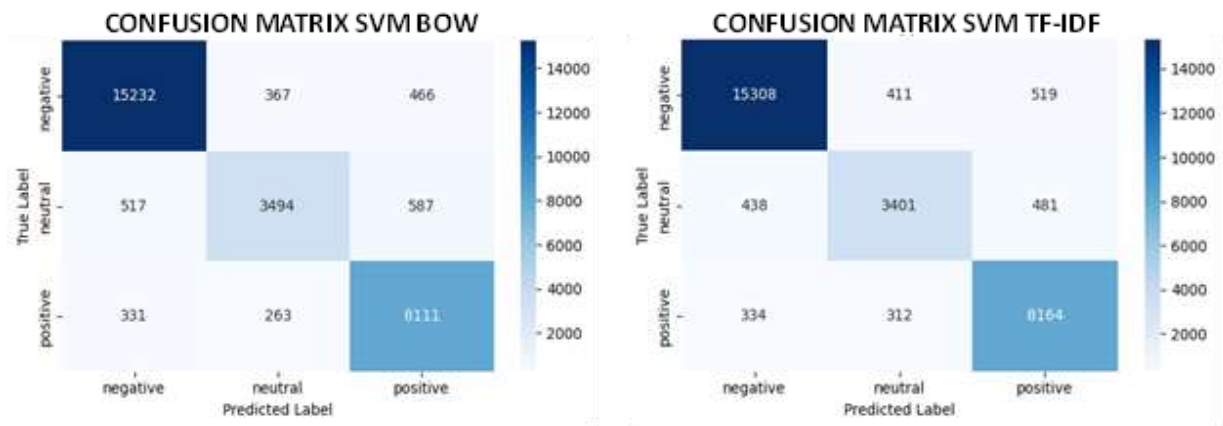
Model	Vektorisasi	Akurasi Pelatihan	Akurasi Pengujian
SVM	TF-IDF	96.44%	87.39%
	BoW	91.65%	91.38%

Pola kinerjanya sangat mirip dengan Logistic Regression. *Classification Report* menunjukkan bahwa SVM sangat andal untuk kelas negatif (*recall* 0.94) dan positif (*recall* 0.93). Namun, model ini juga menunjukkan "kelemahan" yang sama persis: kesulitan dalam mengklasifikasikan ulasan netral, dengan *recall* yang identik yaitu 0.79. Kesamaan pola kesalahan antara SVM dan Logistic Regression ini mengindikasikan bahwa keterbatasan kemungkinan tidak hanya terletak pada algoritma, tetapi juga pada representasi fitur (TF-IDF) yang kesulitan membedakan teks netral yang ambigu. Berikut tabel 10 hasil confusion matrix model SVM:

Tabel 12. Perbandingan Hasil Confusion Matrix Svm

Method	Accuracy		Precision		Recall		F1-Score	
	Train	Test	Macro	Weighted	Macro	Weighted	Macro	Weighted
SVM + TF-IDF	96.44%	87.4%	86.0%	89.00%	87.0%	89.00%	86.5%	89.00%
SVM + BoW	96.46%	87.2%	85.5%	89.00%	86.0%	89.00%	85.7%	89.00%

Gambar 7 hasil model SVM yang dikombinasikan dengan TF-IDF maupun BoW menunjukkan performa tinggi, dengan akurasi masing-masing sebesar 92% dan 91%. Sentimen *negative* dan *positive* diklasifikasikan dengan sangat baik, terbukti dari nilai *precision* dan *recall* di atas 0.93. Namun, tantangan tetap terlihat pada sentimen *neutral* dengan *recall* sebesar 0.79 (TF-IDF) dan 0.76 (BoW), yang menunjukkan bahwa model masih kesulitan mengenali kelas netral secara konsisten. Meski demikian, nilai *macro avg* dan *weighted avg* f1-score tetap tinggi, yaitu 0.89 dan 0.91, yang mencerminkan kinerja model yang solid dengan sedikit keunggulan pada penggunaan TF-IDF.



Gambar 7. Confusion Matrix SVM

Analisis Keseluruhan Model Machine Learning

Kedelapan kombinasi model ML dan vektorisasi dievaluasi. Tabel berikut merangkum akurasi yang dicapai pada data uji (*test accuracy*).

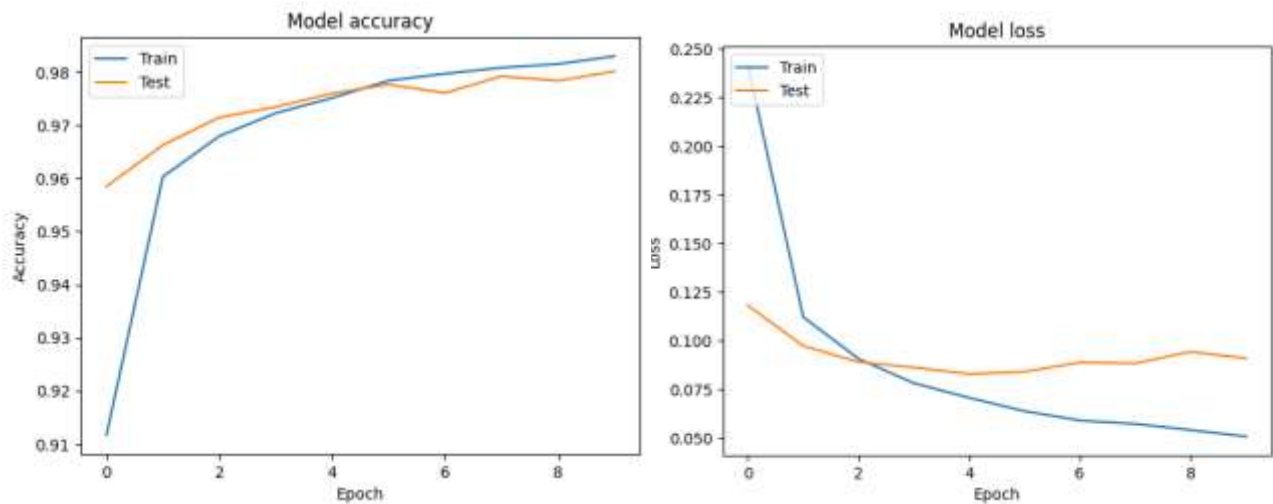
Tabel 12. Perbandingan Kinerja Model ML Berdasarkan Vektorisasi Tf-Idf Dan Bow

Model	Vectorization	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	Notes
Random Forest	TF-IDF	96.44%	90.23%	0.90	0.90	0.90	High on negative/positive
Logistic Regression	TF-IDF	91.66%	91.56%	0.92	0.92	0.92	Most balanced overall
Decision Tree	TF-IDF	96.44%	87.39%	0.87	0.87	0.87	Tends to overfit
SVM (Linear)	TF-IDF	96.44%	87.39%	0.91	0.92	0.91	Great on precision
Random Forest	BoW	96.46%	89.86%	0.90	0.90	0.90	Stable like TF-IDF
Logistic Regression	BoW	91.54%	91.27%	0.91	0.91	0.91	Very competitive
Decision Tree	BoW	96.46%	87.24%	0.87	0.87	0.87	Same issue as with TF-IDF
SVM	BoW	91.65%	91.38%	0.91	0.91	0.91	Best recall on positive class

Dari hasil di atas, Logistic Regression dengan fitur TF-IDF menunjukkan kinerja terbaik di antara model-model ML, dengan akurasi pengujian sebesar 91.56% dan F1-Score sebesar 92%. Model ini juga menunjukkan selisih terkecil antara akurasi pelatihan dan pengujian, mengindikasikan model yang paling seimbang dan tidak *overfitting* dibandingkan yang lain.

Hasil Eksperimen Deep Learning

Model LSTM, setelah melalui pelatihan dan validasi 10-fold Cross-Validation, dengan metrik akurasi dan *loss* selama sepuluh *epoch*. Kurva akurasi pelatihan dan pengujian yang ditampilkan pada Gambar 8 menunjukkan peningkatan yang konsisten, dengan akurasi validasi mencapai 98.02%. Secara bersamaan, kurva *loss* pelatihan dan pengujian pada Gambar 8 memperlihatkan penurunan yang stabil, menandakan konvergensi yang efektif dan *overfitting* yang minimal. Kurva-kurva ini menggambarkan lintasan pembelajaran model serta kemampuannya untuk melakukan generalisasi secara efektif terhadap data yang belum pernah dilihat sebelumnya.



Gambar 8. Training and Testing accuracy and validation loss curve

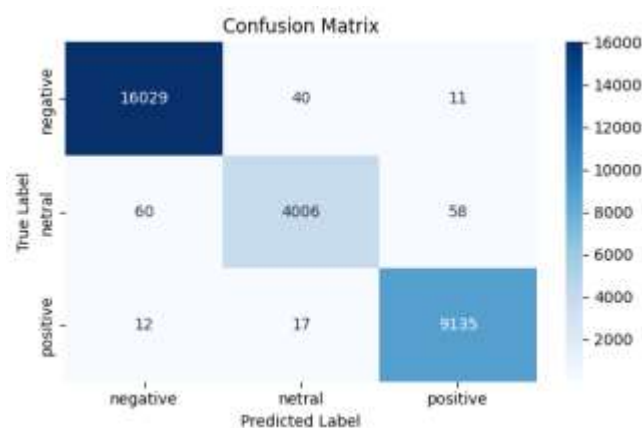
Confusion matrix untuk model LSTM menunjukkan diagonal yang nyaris sempurna, menandakan kesalahan klasifikasi yang sangat minimal. Untuk kelas 'netral', dari 4.124 ulasan, model ini berhasil mengidentifikasi sekitar 3.990 ulasan dengan benar (*recall* 97%). Tingkat kesalahan pada kelas 'negatif' dan 'positif' bahkan lebih rendah lagi, dengan presisi dan *recall* mendekati 1.00. Ini membuktikan bahwa model LSTM tidak hanya akurat secara umum, tetapi juga sangat tangguh dalam membedakan kelas-kelas yang sulit sekalipun.

Tabel 13. Perbandingan Hasil Akurasi Dan Confusion Matrix Lstm

Method	Accuracy	Precision		Recall		F1-Score	
		Macro	Weighted	Macro	Weighted	Macro	Weighted
LSTM	98.02%	99.33%	99.02%	99.00%	99.02%	99.00%	99.02%

Gambar 9 menampilkan confusion matrix yang secara visual menguatkan temuan ini dengan menyoroti jumlah prediksi yang akurat secara signifikan pada diagonal. Kebingungan antara kelas netral dan negatif terlihat jelas, disertai dengan lemahnya ingatan terhadap sentimen netral. Hal ini sesuai dengan nuansa bahasa yang membedakan antara istilah netral dan sedikit negatif dalam ulasan pengguna.

Gambar 9. Confusion Matrix LSTM



Analisis Komparatif

Analisis komparatif dari berbagai model menunjukkan perbedaan yang signifikan dalam kinerja antara machine learning (ML) klasik dan deep learning (DL), khususnya model LSTM. Model

Random Forest, baik dengan vektorisasi TF-IDF (akurasinya 90,23%) maupun BoW (akurasinya 89,86%), menunjukkan hasil yang stabil, tetapi cenderung lebih unggul dalam mengklasifikasikan sentimen negatif dan positif, meskipun ada indikasi overfitting pada data pelatihan. Logistic Regression dengan TF-IDF (akurasi 91,56%) menunjukkan kinerja paling seimbang, dengan precision dan recall yang sangat baik untuk semua kelas, terutama untuk sentimen positif dan negatif. Sebaliknya, model Decision Tree baik dengan TF-IDF maupun BoW menunjukkan kinerja terendah dengan akurasi sekitar 87%, dan mengalami masalah overfitting yang signifikan, meskipun cukup baik dalam mengklasifikasikan sentimen negatif. SVM dengan TF-IDF dan BoW memiliki precision yang sangat baik, terutama dalam mengidentifikasi sentimen negatif dan positif, tetapi cenderung kesulitan dalam menangani kelas netral, meskipun memberikan recall yang baik untuk kelas positif. Di sisi lain, model LSTM dengan akurasi pengujian 98,02% dan weighted F1-score 0,99 menunjukkan kinerja yang sangat superior dibandingkan dengan model-model ML lainnya, dengan kemampuan luar biasa dalam mengidentifikasi nuansa linguistik dan konteks sentimen yang lebih kompleks. Dengan precision dan recall hampir mendekati 1,00 untuk semua kelas, LSTM mengatasi tantangan dalam analisis sentimen dengan kemampuan untuk memahami dependensi sekuensial dalam teks, yang tidak dimiliki oleh model-model berbasis TF-IDF atau BoW. Secara keseluruhan, LSTM tidak hanya mengungguli model-model ML klasik dalam hal akurasi dan kinerja, tetapi juga menunjukkan kemampuannya dalam menangani data yang lebih kompleks dan tidak terstruktur dengan cara yang lebih efisien dan akurat.

Perbandingan langsung antara model ML terbaik dengan model DL menunjukkan perbedaan kinerja yang signifikan. Hasil ini dapat dilihat jelas pada tabel 14 yang menunjukkan bahwa arsitektur LSTM **secara signifikan mengungguli** pendekatan ML klasik dengan selisih performa F1-Score sebesar **7%**.

Tabel 14. Hasil Akurasi Dan Confusion Matrix Model Terbaik

Model Terbaik	Akurasi Pengujian	Weighted F1-Score
Logistic Regression + TF-IDF	91.56%	0.92
Model LSTM	98.02%	0.99

PEMBAHASAN

Keunggulan signifikan model *Long Short-Term Memory* (LSTM) dalam penelitian ini dapat dijelaskan melalui kapabilitasnya dalam memahami konteks dan dependensi sekuensial pada data teks. Berbeda dengan pendekatan tradisional seperti *Bag-of-Words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang memperlakukan teks sebagai "kantong kata" tanpa mempertimbangkan urutan, LSTM memproses kalimat secara sekuensial. Kemampuan ini memungkinkan LSTM untuk menangkap nuansa halus dalam bahasa alami, termasuk ekspresi idiomatik dan negasi seperti "tidak puas." Temuan ini diperkuat oleh studi (Alawaji & Aloraini, 2025), yang menunjukkan bahwa LSTM dan *Large Language Models* secara konsisten mengungguli pendekatan klasik dalam menganalisis ulasan pengguna perbankan digital berkat kemampuannya dalam memahami konteks bahasa informal dan tidak terstruktur. Hal ini juga didukung oleh penelitian oleh (Du et al., 2021), yang menyimpulkan bahwa model LSTM berbasis leksikon mampu mengklasifikasikan ulasan dengan akurasi tinggi, bahkan dengan data masukan yang tidak terstruktur.

Secara teoretis dan praktis, temuan penelitian ini menawarkan implikasi yang signifikan. Dalam konteks industri FinTech, LSTM terbukti menjadi sistem pemantauan sentimen otomatis yang andal, yang memungkinkan institusi keuangan, dalam hal ini Bank Mandiri, untuk merespons umpan balik pelanggan secara lebih proaktif. Studi lain menemukan bahwa meskipun teknik *ensemble learning* mencapai akurasi tinggi dalam domain perbankan digital Arab, pendekatan *deep learning* seperti LSTM memiliki keunggulan dalam menangani dimensi semantik dan nuansa lokal (Ashbaugh & Zhang, 2024). Hal ini mempertegas bahwa dalam konteks Indonesia yang kaya akan keragaman linguistik, termasuk maraknya penggunaan bahasa gaul (*slang*), singkatan, dan campur kode (*code-switching*), pendekatan *deep learning* seperti LSTM memberikan keunggulan fundamental dibandingkan model klasik. Dukungan lebih lanjut datang dari (Sharma & Goyal, 2023), yang menunjukkan bahwa

model Bi-LSTM dapat mencapai akurasi di atas 90% saat mengklasifikasikan komentar informal di platform digital. Ini mengindikasikan efektivitas model berbasis memori dalam menangani kalimat yang panjang dan kompleks secara kontekstual.

Namun demikian, pendekatan ini memiliki keterbatasan, terutama pada proses pelabelan berbasis leksikon yang kurang adaptif terhadap ekspresi non-literal seperti sarkasme. Oleh karena itu, penelitian di masa depan dapat diarahkan untuk mengeksplorasi metode *semi-supervised* atau melakukan *fine-tuning* pada model bahasa yang spesifik untuk konteks Indonesia, seperti IndoBERT. Arah ini sejalan dengan temuan (Zakarija et al., 2024), yang menunjukkan bahwa model ABSA berbasis IndoBERT mampu meningkatkan akurasi ekstraksi aspek sentimen hingga 97%, terutama pada domain *e-commerce* Indonesia yang juga kaya akan bahasa informal. Selain itu, perluasan cakupan penelitian ke aplikasi digital lainnya akan memperkuat generalisasi temuan ini. Secara metodologis, penerapan *Aspect-Based Sentiment Analysis* (ABSA) dapat memberikan wawasan yang lebih granular dan dapat ditindaklanjuti (*actionable*). Pendekatan ini memungkinkan identifikasi sentimen pada level fitur layanan yang spesifik, seperti 'fitur transfer', 'tampilan UI/UX', atau 'responsivitas layanan pelanggan', sehingga memberikan nilai praktis yang lebih tinggi bagi pemangku kepentingan.

KESIMPULAN

Penelitian ini secara komprehensif mengevaluasi kinerja berbagai model machine learning (ML) dan deep learning (DL) dalam analisis sentimen pada ulasan pengguna Livin' by Mandiri, sebuah super-app perbankan digital di Indonesia. Temuan utama menunjukkan bahwa model Long Short-Term Memory (LSTM), dengan kemampuannya dalam menangkap dependensi sekuensial dan konteks jangka panjang dalam teks, secara signifikan mengungguli model ML klasik seperti Logistic Regression, Random Forest, Decision Tree, dan Support Vector Machine (SVM). LSTM mencapai akurasi pengujian 98,02% dan weighted F1-score 0,99, sementara model ML terbaik hanya mencapai akurasi sekitar 91%. Keunggulan LSTM terletak pada kemampuannya untuk memahami nuansa linguistik dalam bahasa informal, yang sering ditemui dalam ulasan pengguna, dan memproses hubungan kontekstual yang kompleks, seperti negasi dan frasa idiomatik. Implikasi praktis dari temuan ini sangat besar bagi industri FinTech, khususnya dalam pengembangan sistem pemantauan sentimen otomatis yang dapat meningkatkan responsivitas terhadap umpan balik pengguna secara real-time dan mengoptimalkan pengalaman pelanggan. Secara teoretis, penelitian ini memperkuat bukti bahwa untuk domain dengan kompleksitas linguistik tinggi, seperti Bahasa Indonesia, model DL menawarkan performa yang jauh lebih superior dibandingkan model ML klasik. Keterbatasan penelitian ini, seperti penggunaan pelabelan berbasis leksikon yang terbatas dalam menangani slang atau sarkasme, membuka peluang untuk pengembangan lebih lanjut dengan menggunakan pendekatan semi-supervised learning atau fine-tuning model bahasa canggih, seperti IndoBERT. Secara keseluruhan, hasil penelitian ini memberikan kontribusi signifikan baik untuk pengembangan teknologi analisis sentimen di industri perbankan digital maupun untuk penelitian lebih lanjut dalam domain pemrosesan bahasa alami yang lebih kompleks.

DAFTAR PUSTAKA

- Alawaji, R., & Aloraini, A. (2025). Sentiment Analysis of Digital Banking Reviews Using Machine Learning and Large Language Models. *Electronics (Switzerland)*, 14(11). <https://doi.org/10.3390/electronics14112125>
- Alqahtani, A., Khan, S. B., Alqahtani, J., AlYami, S., & Alfayez, F. (2023). Sentiment Analysis of Semantically Interoperable Social Media Platforms Using Computational Intelligence Techniques. *Applied Sciences (Switzerland)*, 13(13). <https://doi.org/10.3390/app13137599>
- Angel Metanosa Afinda. (n.d.). <https://github.com/angelmetanosaa/dataset> (Accessed on Juli 14, 2024).
- Ashbaugh, L., & Zhang, Y. (2024). A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning. *Computers*.
- Du, M., Li, X., & Luo, L. (2021). A Training-Optimization-Based Method for Constructing Domain-Specific Sentiment Lexicon. In *Complexity* (Vol. 2021). Hindawi Limited. <https://doi.org/10.1155/2021/6152494>

- Ependi, U., Muzakir, A., & Wibowo, A. (2023). Sentiment Analysis on Smart City Mobile Platform Based on Lexicon. *IEEE International Conference on Smart Technology*.
- Garg, N., & Sharma, K. (2022). Text pre-processing of multilingual for sentiment analysis based on social network data. *International Journal of Electrical and Computer Engineering*, 12(1), 776–784. <https://doi.org/10.11591/ijece.v12i1.pp776-784>
- Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., & Shaukat, N. (2024). Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM. *Big Data and Cognitive Computing*, 8(12). <https://doi.org/10.3390/bdcc8120199>
- Jazuli, A., Widowati, & Kusumaningrum, R. (2025). Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback. *Applied Sciences (Switzerland)*, 15(1). <https://doi.org/10.3390/app15010172>
- Kavitha, M., Naib, B. B., Mallikarjuna, B., Kavitha, R., & Srinivasan, R. (2022). Sentiment Analysis using NLP and Machine Learning Techniques on Social Media Data. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022*, 112–115. <https://doi.org/10.1109/ICACITE53722.2022.9823708>
- Kim, J., & Chung, J. (2024). Analysis of Service Quality in Smart Running Applications Using Big Data Text Mining Techniques. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(4), 3352–3369. <https://doi.org/10.3390/jtaer19040162>
- Ma'aly, A. N., Pramesti, D., Fathurahman, A. D., & Fakhurroja, H. (2024). Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm. *Information (Switzerland)*, 15(11). <https://doi.org/10.3390/info15110705>
- Nip, J. Y. M., & Berthelie, B. (2024). Social Media Sentiment Analysis. *Encyclopedia*, 4(4), 1590–1598. <https://doi.org/10.3390/encyclopedia4040104>
- Refianti, R., Mutiara, A. B., & Putra, R. A. (2024). A Lexicon-Based Long Short-Term Memory (LSTM) Model for Sentiment Analysis to Classify Halodoc Application Reviews on Google Playstore. *Journal of Applied Data Sciences*, 5(1), 146–157. <https://doi.org/10.47738/jads.v5i1.160>
- Sharma, H. D., & Goyal, P. (2023). An Analysis of Sentiment: Methods, Applications, and Challenges †. *Engineering Proceedings*, 59(1). <https://doi.org/10.3390/engproc2023059068>
- Sindhu, S., Kumar, S., & Noliya, A. (2023). A Review on Sentiment Analysis using Machine Learning. *International Conference on Innovative Data Communication Technologies and Application, ICIDCA 2023 - Proceedings*, 138–142. <https://doi.org/10.1109/ICIDCA56705.2023.10099665>
- Sushma, S., Nayak, S. K., & Krishna, M. V. (2024). A Comprehensive Review of Sentiment Analysis: Trends, Challenges, and Future Directions. *Proceedings of the 5th International Conference on Data Intelligence and Cognitive Informatics, ICDICI 2024*, 1175–1181. <https://doi.org/10.1109/ICDICI62993.2024.10810919>
- Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. In *Applied Sciences (Switzerland)* (Vol. 13, Issue 7). MDPI. <https://doi.org/10.3390/app13074550>
- Torres, E. C. M., & de Picado-Santos, L. G. (2025). Sentiment Analysis and Topic Modeling in Transportation: A Literature Review. In *Applied Sciences (Switzerland)* (Vol. 15, Issue 12). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/app15126576>
- Yang, Q. (2023). Research on E-commerce Customer Satisfaction Evaluation Method Based on PSO-LSTM and Text Mining. *3C Empresa. Investigación y Pensamiento Crítico*, 12(01), 51–66. <https://doi.org/10.17993/3cemp.2023.120151.51-66>
- Zakarija, I., Škopljanač-Maćina, F., Marušić, H., & Blašković, B. (2024). A Sentiment Analysis Model Based on User Experiences of Dubrovnik on the Tripadvisor Platform. *Applied Sciences (Switzerland)*, 14(18). <https://doi.org/10.3390/app14188304>