Indo Green Journal
https://idngreen.com/index.php/green

e-ISSN: 3025 – 4574, p-ISSN: 3025 – 2679
Volume 4 Nomor 4 Tahun 2026
Halaman: 458 – 463

RESEARCH ARTICLE

Analisis Sentimen dan Preprocessing Data Ulasan Aplikasi Access by KAI Menggunakan Metode Scraping dan Natural Language Processing

Ellestiana Muhimmatul Rachmawati[1], Tasya Crisanti Manalu[2], Siti Farianti[3], Yayang Maulani Indyatmaja[4]
Program Studi Pendidikan Teknologi Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Indonesia

ellestiana.23010@mhs.unesa.ac.id[1], tasya.23153@mhs.unesa.ac.id[2], sitifarianti.23180@mhs.unesa.ac.id[3], yayang.23183@mhs.unesa.ac.id[4]

Kata Kunci:
Analisis Sentimen, Access by KAI, Google Play Store, Preprocessing Teks, Natural Language Processing

Abstrak

Aplikasi *Access by KAI* merupakan platform digital utama yang disediakan oleh PT Kereta Api Indonesia (Persero) untuk memfasilitasi pemesanan tiket dan layanan transportasi kereta api di Indonesia. Evaluasi terhadap kepuasan pengguna sangat penting untuk meningkatkan kualitas layanan aplikasi tersebut. Penelitian ini bertujuan untuk menganalisis sentimen pengguna berdasarkan ulasan publik di Google Play Store dengan menerapkan alur kerja pemrosesan bahasa alami (*Natural Language Processing* / NLP) secara *end-to-end*. Data ulasan berhasil diambil menggunakan teknik *scraping* melalui pustaka google-play-scraper. Setelah tahap pembersihan data (*data cleaning*) dan penghapusan ulasan duplikat, diperoleh dataset ulasan bersih. Pelabelan sentimen berbasis tingkatan rating bintang (*rating-based labeling*) menghasilkan distribusi sentimen yang dominan pada kategori negatif. Tahap *preprocessing* dilakukan meliputi *case folding*, *slangword normalization*, *stopword removal* berbasis pembatasan kata negasi, serta *tokenization*. Hasil analisis visualisasi *WordCloud* mengidentifikasi bahwa kendala utama pengguna didominasi oleh isu kegagalan sistem pembayaran, kendala fitur pembatalan tiket, serta ketidakstabilan aplikasi (*app crash* dan *error* saat *login*). Hasil dari penelitian ini memberikan pemahaman terstruktur bagi pengembang untuk memprioritaskan perbaikan fitur kritis pada aplikasi Access by KAI.

Keyword:
Sentiment Analysis, Access by KAI, Google Play Store, Text Preprocessing, Natural Language Processing

Abstract

The Access by KAI application is the primary digital platform provided by PT Kereta Api Indonesia (Persero) to facilitate ticket booking and railway transportation services in Indonesia. Evaluating user satisfaction is crucial to improving the service quality of the application. This study aims to analyze user sentiment based on public reviews on the Google Play Store by implementing an end-to-end Natural Language Processing (NLP) pipeline. Reviews were extracted using web scraping techniques via the google-play-scraper library. Following data cleaning and duplicate removal, a clean dataset was obtained. Sentiment labeling based on rating scores revealed a sentiment distribution dominated by negative feedback. The text preprocessing pipeline included case folding, slangword normalization, custom stopword removal preserving negation words, and tokenization. Visual analysis using WordCloud identified that major user complaints were driven by payment gateway failures, ticket cancellation issues, and application instability (app crashes and login errors). The findings of this research provide structured insights for developers to prioritize critical feature enhancements in the Access by KAI platform.

Information Production:
© 2026 *by the authors of this article. This work is licensed under an Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)*

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mengubah pola interaksi serta transaksi masyarakat dalam memanfaatkan moda transportasi publik. PT Kereta Api Indonesia (Persero) sebagai penyedia jasa transportasi kereta api utama di Indonesia meluncurkan aplikasi *Access by KAI* (sebelumnya dikenal sebagai KAI Access) guna memberikan kemudahan aksesibilitas bagi penumpang dalam melakukan pemesanan tiket, pengecekan jadwal, hingga layanan pembatalan atau perubahan jadwal secara digital. Seiring dengan peningkatan jumlah pengguna yang telah menembus lebih dari 10 juta unduhan pada platform Google Play Store, ulasan yang ditinggalkan oleh pengguna menjadi sumber data yang sangat berharga untuk mengevaluasi kualitas layanan dan performa perangkat lunak.

Ulasan pengguna (*user reviews*) mengekspresikan pengalaman langsung, kepuasan, maupun kendala teknis yang dialami selama menggunakan aplikasi. Namun, volume data ulasan yang sangat besar dan bersifat tidak terstruktur (*unstructured data*) menyulitkan proses evaluasi secara manual. Oleh karena itu, pendekatan *Natural Language Processing* (NLP) dan analisis sentimen menjadi solusi efektif untuk mengekstraksi

e-ISSN: 3025 – 4574
p-ISSN: 3025 – 2679

informasi serta mengelompokkan persepsi publik secara otomatis ke dalam kategori positif, netral, maupun negatif (Bahtiar et al., 2023; Lestari & Hutagalung, 2025).

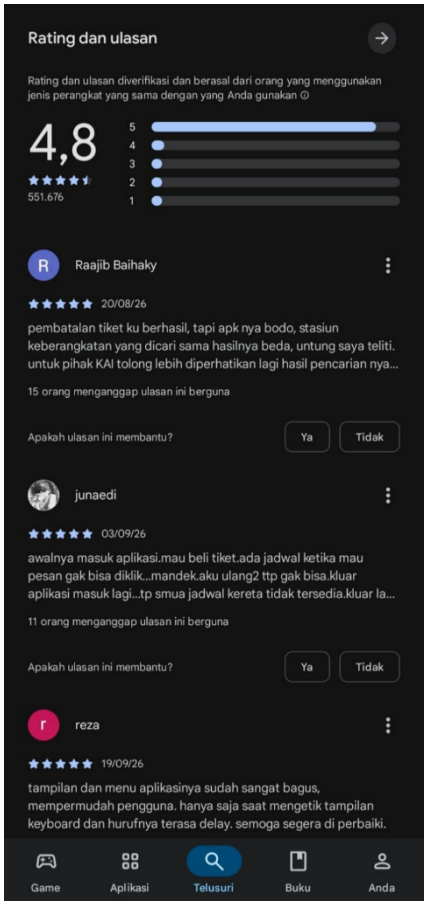
Beberapa penelitian terdahulu telah menyoroti pentingnya analisis sentimen dan pemrosesan teks pada aplikasi KAI Access maupun platform digital publik lainnya. Damanhuri dan Husein (2024) serta Sholiihah et al. (2024) melakukan analisis ulasan KAI Access berbasis pembelajaran mesin (*machine learning*) dan menemukan bahwa aspek teknis sistem menjadi fokus utama persepsi pengguna. Di sisi lain, evaluasi aspek spesifik juga dikembangkan melalui pendekatan *aspect-based sentiment analysis* untuk mengenali keluhan fungsional secara terperinci (Mustakim & Priyanta, 2022). Selain model pemodelan, efektivitas prapemrosesan teks (*text preprocessing*) seperti normalisasi kata slang, *case folding*, dan ekstraksi fitur memegang peranan sangat krusial dalam meningkatkan kualitas data teks sebelum dianalisis lebih lanjut (Chai, 2023; Putri et al., 2026).

Meskipun demikian, sebagian besar penelitian sebelumnya terbatas pada sampel data berukuran relatif kecil atau berfokus pada pemodelan klasifikasi. Penelitian ini bertujuan untuk melakukan penambangan data (*data mining*) ulasan aplikasi *Access by KAI* dalam skala besar dengan menerapkan alur *preprocessing* teks bahasa Indonesia yang terintegrasi. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi empiris dalam pemrosesan teks ulasan aplikasi layanan publik serta memberikan rekomendasi praktis bagi pengembang *Access by KAI* dalam meningkatkan kualitas pengalaman pengguna (*user experience*).

METODE PENELITIAN

Metodologi yang digunakan dalam penelitian ini mengikuti tahapan alur kerja pemrosesan data (*data processing pipeline*) berbasis Python. Penelitian dilakukan melalui lima tahapan utama, yaitu: *data scraping, data cleaning, text preprocessing, labeling, dan Exploratory Data Analysis* (EDA) serta klasifikasi sentimen.

Data ulasan diperoleh secara otomatis dari platform Google Play Store menggunakan pustaka google-play-scraper. Pengambilan data difokuskan pada aplikasi *Access by KAI* dengan ID paket com.kai.kaiticketing untuk mengumpulkan ulasan terbaru berbahasa Indonesia. Atribut data yang diambil meliputi nama pengguna, isi ulasan (*content*), jumlah bintang/rating (*score*), serta waktu pengiriman ulasan.



Gambar 1. Ulasan Pengguna Aplikasi Access by KAI di Google Play Store

e-ISSN: 3025 – 4574
p-ISSN: 3025 – 2679

```
# Cek dimensi dataset (Baris, Kolom)
print(f"Jumlah baris awal  : {df.shape[0]}")
print(f"Jumlah kolom awal : {df.shape[1]}")
print("-" * 40)

# Cek tipe data dan nama-nama kolom
df.info()
```

```
Jumlah baris awal  : 100000
Jumlah kolom awal : 11
----------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 11 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   reviewId              100000 non-null  object
 1   userName              100000 non-null  object
 2   userImage             100000 non-null  object
 3   content               100000 non-null  object
 4   score                 100000 non-null  int64
 5   thumbsUpCount         100000 non-null  int64
 6   reviewCreatedVersion  81403 non-null   object
 7   at                    100000 non-null  datetime64[ns]
 8   replyContent          11587 non-null   object
 9   repliedAt             11587 non-null   datetime64[ns]
 10  appVersion            81403 non-null   object
dtypes: datetime64[ns](2), int64(2), object(7)
memory usage: 8.4+ MB
```

Gambar 2. Output Tampilan Struktur Data Awal Dataset Ulasan

Tahap awal pembersihan data dilakukan untuk menjamin validitas dataset dengan menghapus baris data yang mengalami *missing values* serta data duplikat berdasarkan kesamaan isi teks ulasan.

Dataset teks ulasan mengalami beberapa tahap konversi dan pembersihan menggunakan pustaka Sastrawi, NLTK, dan ekspresi reguler (re) dengan mengacu pada teknik prapemrosesan teks bahasa Indonesia (Chai, 2023; Putri et al., 2026): a). Case Folding & Cleansing: Mengubah seluruh teks menjadi huruf kecil (*lowercase*), menghapus URL, *mention* (@), *hashtag* (#), angka, tanda baca, serta karakter non-ASCII (seperti emotikon). b). Normalisasi Slangword: Mengganti kata-kata singkatan atau bahasa gaul menjadi kata baku sesuai kamus istilah (misalnya kata "apk" disesuaikan menjadi "aplikasi", "gak" menjadi "tidak"). c). Stopwords Removal: Menghapus kata-kata umum yang tidak memuat informasi sentimen spesifik, dengan mempertahankan kata negasi penting (seperti *tidak*, *kurang*, *jangan*, *bukan*). d). Tokenisasi (Tokenization): Memecah kalimat ulasan yang telah bersih menjadi potongan kata (*tokens*).

	normalized_text	stopwords_text
0	oke	oke
1	pembatalan tiket gagal terus padahal kode veri...	pembatalan tiket gagal kode verifikasi udah ul...
2	oke	oke
3	ini untuk perubahan jadwal atau pembatalan kok...	perubahan jadwal pembatalan sulit diklik tidak...
4	baik	

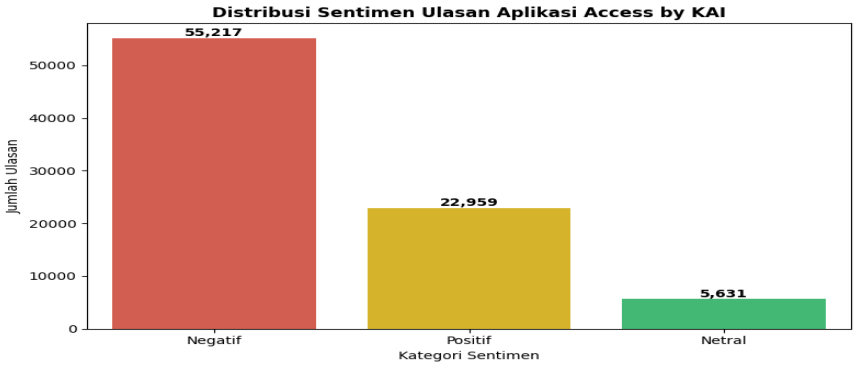
Gambar 3. Perbandingan Output Teks Raw Text vs Clean Text

Setelah tahap prapemrosesan teks selesai, pelabelan sentimen dilakukan secara otomatis berbasis nilai rating (rating-based labeling) sesuai pendekatan yang diterapkan oleh Bahtiar et al. (2023), di mana rating 1 & 2 dikategorikan sebagai Negatif, rating 3 sebagai Netral, serta rating 4 & 5 sebagai Positif.

HASIL DAN PEMBAHASAN

Proses pembersihan data berhasil menghapus duplikasi dan missing value. Pelabelan sentimen otomatis berdasarkan skala nilai rating menghasilkan 83.807 data ulasan bersih. Berdasarkan pengolahan dataset tersebut, diperoleh distribusi kategori sentimen yang didominasi oleh kategori Negatif sebanyak 55.217 ulasan (65,89%), diikuti oleh kategori Positif sebesar 22.959 ulasan (27,40%), serta kategori Netral sebanyak 5.631 ulasan (6,72%) seperti yang disajikan pada Gambar 4.

e-ISSN: 3025 – 4574
p-ISSN: 3025 – 2679

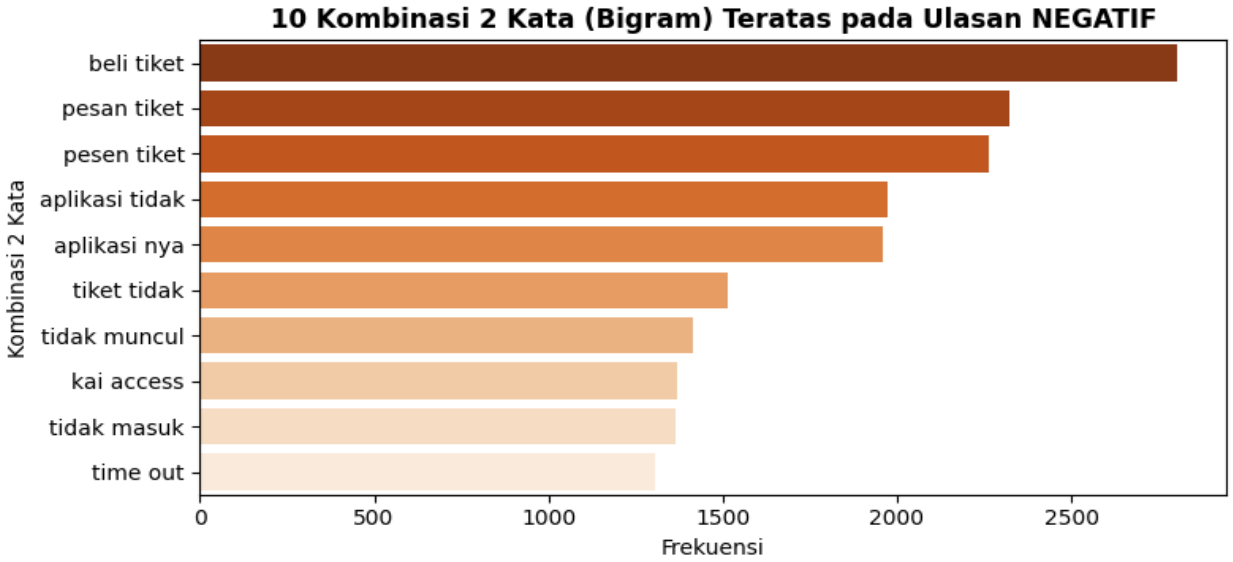


Gambar 4. Grafik Batang Distribusi Label Sentimen

Berdasarkan hasil visualisasi kata kunci menggunakan *WordCloud* dan analisis Bigram setelah pemrosesan *slangword* dan *stopwords removal*, teridentifikasi sebaran kata kunci utama pengguna yang ditampilkan pada Gambar 5 dan Gambar 6.



Gambar 5. Visualisasi WordCloud Sentimen Negatif (Merah) dan Positif (Hijau)



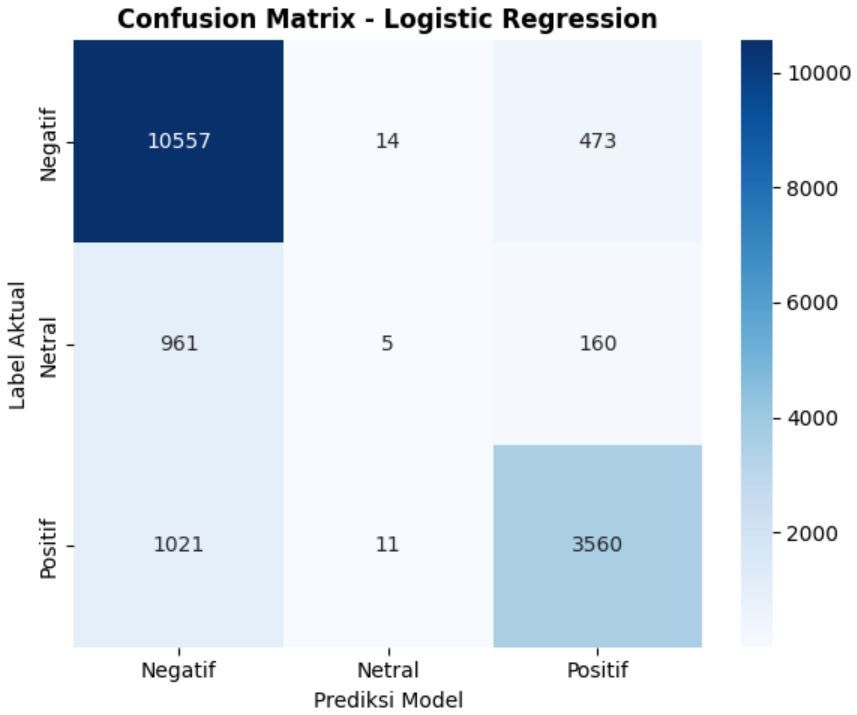
Gambar 6. Grafik Batang 10 Kombinasi Kata (Bigram) Teratas Ulasan Negatif

Pengujian klasifikasi sentimen dilakukan menggunakan algoritma pembelajaran mesin seperti Naïve Bayes, Logistic Regression, dan Support Vector Machine (SVM). Hasil evaluasi performa model klasifikasi serta matriks konfusi (*confusion matrix*) disajikan pada Gambar 7 dan Gambar 8.

```
=== TABEL PERBANDINGAN PERFORMA MODEL ===
                          Model  Akurasi (%)  Precision (%)  Recall (%)  F1-Score (%)
        Multinomial Naive Bayes        81.76          76.92       81.76         78.34
            Logistic Regression        84.25          79.85       84.25         81.25
Support Vector Machine (SVM)           83.68          79.12       83.68         80.84
```

Gambar 7. Tabel Perbandingan Evaluasi Performa Model Klasifikasi

e-ISSN: 3025 – 4574
p-ISSN: 3025 – 2679



Gambar 8. Confusion Matrix Heatmap Model Logistic Regression

Pembahasan

Data pada Gambar 4 menunjukkan bahwa mayoritas ulasan pengguna bersifat negatif sebanyak 55.217 ulasan (65,89%), sedangkan ulasan positif mencakup 22.959 ulasan (27,40%), dan sisanya ulasan netral sebesar 5.631 ulasan (6,72%). Tingginya proporsi sentimen negatif ini sejalan dengan temuan Damanhuri dan Husein (2024) serta Sholiihah et al. (2024), di mana perubahan sistem atau pembaruan antarmuka aplikasi sering kali memicu peningkatan keluhan pengguna terkait stabilitas sistem.

Kata kunci yang dominan muncul pada kelompok ulasan negatif meliputi kata-kata seperti *"gagal"*, *"pembayaran"*, *"batal"*, *"e-wallet"*, *"error"*, *"login"*, *"aplikasi"*, dan *"berat"*. Temuan ini mengindikasikan bahwa kendala terbesar yang dihadapi pengguna tidak bersumber pada layanan fisik kereta api, melainkan pada sistem integrasi pembayaran digital serta stabilitas teknis aplikasi saat memuat halaman (*loading/crash*) atau proses pembaruan sistem (*update*), yang juga dikonfirmasi dalam penelitian berbasis aspek oleh Mustakim dan Priyanta (2022).

Sebaliknya, pada kelompok ulasan positif kata kunci yang muncul didominasi oleh kata *"mudah"*, *"membantu"*, *"bagus"*, *"lancar"*, *"pesan"*, dan *"tiket"*. Hal ini mengindikasikan bahwa pengguna memberikan apresiasi positif terhadap efisiensi dan kemudahan akses pemesanan tiket secara mandiri tanpa harus mengantre di stasiun ketika sistem aplikasi berjalan normal.

KESIMPULAN

Penelitian ini berhasil membangun dan mengimplementasikan alur kerja analisis sentimen berbasis *Natural Language Processing* (NLP) untuk mengolah ulasan aplikasi *Access by KAI* dari Google Play Store. Hasil analisis menunjukkan dominasi sentimen negatif sebesar 65,89%, diikuti sentimen positif 27,40%, dan netral 6,72%. Kendala utama yang dikeluhkan oleh pengguna berkisar pada masalah kegagalan transaksi pembayaran, proses pembatalan tiket, serta ketidakstabilan aplikasi (*system crash/error login*). Disarankan bagi pengembang aplikasi *Access by KAI* untuk memprioritaskan optimalisasi performa *backend server*, perbaikan *payment gateway*, serta penyederhanaan alur pembatalan tiket demi meningkatkan kepuasan pengguna.

DAFTAR REFERENSI

Bahtiar, S. A. H., Dewa, C. K., & Luthfi, A. (2023). Comparison of Naïve Bayes and logistic regression in sentiment analysis on marketplace reviews using rating-based labeling. *Journal of Information Systems and Informatics, 5*(3), 915–927. https://doi.org/10.51519/journalisi.v5i3.539

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering, 29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

e-ISSN: 3025 – 4574
p-ISSN: 3025 – 2679

Damanhuri, R., & Husein, V. A. (2024). Analisis sentimen pada ulasan aplikasi Access by KAI berbahasa Indonesia menggunakan word-embedding dan classical machine learning. *Jurnal Masyarakat Informatika, 15*(2), 97–106. https://doi.org/10.14710/jmasif.15.2.62383

Lestari, V. B., & Hutagalung, C. A. (2025). Evaluation of TF-IDF extraction techniques in sentiment analysis of Indonesian-language marketplaces using SVM, logistic regression, and naive Bayes. *J-KOMA: Jurnal Ilmu Komputer dan Aplikasi, 8*(1), 36–44. https://doi.org/10.21009/j-koma.v8i1.05

Mustakim, H., & Priyanta, S. (2022). Aspect-based sentiment analysis of KAI Access reviews using NBC and SVM. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 16*(2), 113–124. https://doi.org/10.22146/ijccs.68903

Putri, J. R., Wijaya, H., Ibrahim, A., & Heryanto, A. (2026). Optimized Indonesia language preprocessing framework for Gojek reviews sentiment analysis using linear support vector machine. *Indonesian Journal of Artificial Intelligence and Data Mining, 9*(2). https://ejournal.uin-suska.ac.id/index.php/IJAIDM/article/view/38793

Sholiihah, E. R. M., Diyasa, I. G. S. M., & Puspaningrum, E. Y. (2024). Perbandingan kinerja kernel linear dan RBF support vector machine untuk analisis sentimen ulasan pengguna KAI Access pada Google Play Store. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(1), 728–733.