JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

Perbandingan Sentimen Ulasan Pengguna Aplikasi Brainly dan Ruangguru Menggunakan Naïve bayes, KNN, Decision Tree

Adrianus Windi[1*], Panny Agustia Rahayuningsih[2], Muhammad Rezki[3]

[1,2,3]Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika, Pontianak, Indonesia

Email: [1*]adrianuswindi227@gmail, [2]muhammad.mdk@bsi.ac.id, [3]Panny.par@bsi.ac.id

(*Email Corresponding Author: adrianuswindi227@gmail)

Received: 18 Juni 2026 | Revision: 21 Juni 2026 | Accepted: 25 Juni 2026

Abstrak

Analisis sentimen adalah cara yang penting untuk memahami pendapat pengguna tentang aplikasi pendidikan digital, karena jumlah ulasan di Google Play Store terlalu banyak untuk dijelaskan secara manual satu per satu. Penelitian ini membandingkan tiga metode pembelajaran mesin, yaitu Naïve Bayes, K-Nearest Neighbor (KNN), dan Decision Tree, untuk mengklasifikasikan perasaan dari ulasan pengguna aplikasi Brainly dan Ruang Guru. Data dikumpulkan dengan cara mengambil dari Google Play Store sebanyak 8.000 ulasan , yaitu 4.000 ulasan per aplikasi , dari bulan Mei hingga Juni 2026; setelah menghilangkan ulasan yang sama, tersisa 6.151 ulasan , terdiri dari 2.836 ulasan untuk Brainly dan 3.315 ulasan untuk Ruang Guru. Label sentimen diatur berdasarkan jumlah bintang (1–3 berarti negatif , 4–5 berarti positif), sehingga menghasilkan distribusi yang tidak seimbang yaitu 79,8% positif dan 20,2% negatif. Teks diproses melalui sembilan tahap pra-pengolahan yang khusus digunakan untuk bahasa Indonesia informal . *Fitur* kemudian diambil menggunakan metode TF-IDF dengan menghasilkan 2.398 fitur dan tingkat penayangan sebesar 99,78%. Data latih disamakan kuantitas menggunakan teknik SMOTE, dan model dioptimalkan dengan *GridSearchCV* yang menggunakan *StratifiedKFold* dengan 5 kali pembagian data . Dalam skenario *tuning* dan SMOTE, *metode Naïve Bayes* menunjukkan performa terbaik dengan akurasi sebesar 82,78%, F1- *Score* mencapai 83,79%, dan ROC-AUC sebesar 88,44%, yang lebih baik dibandingkan *Decision Tree* dan KNN. Menariknya , metode *Naïve Bayes* tanpa menggunakan SMOTE justru mencapai akurasi tertinggi secara keseluruhan, yaitu 88,95%, yang menunjukkan bahwa penggunaan SMOTE pada data TF-IDF berdimensi tinggi tidak selalu meningkatkan kinerja model . *Analisis* kata kunci pembeda membantu mengenali sentimen positif seperti ' membantu ', 'mudah', dan 'terbaik', serta sentimen negatif seperti ' sampah ', 'iklan', dan ' error ', yang bisa digunakan sebagai tolak ukur dalam menyediakan kualitas layanan oleh pengembang aplikasi kedua.

Kata Kunci: Analisis Sentimen, Naïve Bayes, K-Nearest Neighbor, Decision Tree, Google Play Store.

Abstract

Sentiment analysis is an important way to understand user opinions about digital education apps, as the number of reviews on the Google Play Store is too large to be manually analyzed one by one. This study compares three machine learning methods, namely Naïve Bayes, K-Nearest Neighbor (KNN), and Decision Tree, to classify sentiments from user reviews of the Brainly and Ruang Guru apps. Data were collected by scraping 8,000 reviews from the Google Play Store, i.e., 4,000 reviews per app, from May to June 2026; after removing duplicate reviews, 6,151 reviews remained, consisting of 2,836 reviews for Brainly and 3,315 reviews for Ruang Guru. Sentiment labels were arranged based on the number of stars (1–3 means negative, 4–5 means positive), resulting in an unbalanced distribution of 79.8% positive and 20.2% negative. The text was processed through nine pre-processing stages specifically used for informal Indonesian. Features were then extracted using the TF-IDF method, resulting in 2,398 features and a viewing rate of 99.78%. The training data was quantity-equalized using the SMOTE technique, and the model was optimized with GridSearchCV using StratifiedKFold with 5 data splits. In the tuning and SMOTE scenarios, the Naïve Bayes method showed the best performance with an accuracy of 82.78%, an F1-Score of 83.79%, and an ROC-AUC of 88.44%, which was better than Decision Tree and KNN. Interestingly, the Naïve Bayes method without using SMOTE actually achieved the highest overall accuracy of 88.95%, indicating that using SMOTE on high-dimensional TF-IDF data does not always improve model performance. Differentiating keyword analysis helps to identify positive sentiments such as 'helpful', 'easy', and 'best', as well as negative sentiments such as 'trash', 'ads', and 'error', which can be used as a benchmark in providing service quality by the second application developer.

***Keywords**: Sentiment analysis, Naïve Bayes, K-Nearest Neighbor, Decision Tree, Google Play Store.*

1. PENDAHULUAN

Percepatan digitalisasi pendidikan di Indonesia mengubah cara siswa memperoleh materi belajar lewat aplikasi pembelajaran berbasis perangkat bergerak. Brainly dan Ruang Guru termasuk dua platform yang banyak diminati karena menawarkan pendekatan layanan berlainan: Brainly berpusat pada forum tanya-jawab antarpelajar, sedangkan Ruang Guru mengusung sistem belajar terstruktur lewat video materi, latihan soal, dan kelas bimbingan daring. Kolom ulasan di Google Play Store menjadi ruang bagi pengguna untuk menuangkan pengalaman, kepuasan, maupun keluhan mereka. Ulasan ini menyimpan informasi berharga, hanya saja jumlahnya yang banyak menjadikan penelaahan manual sulit dilakukan.

Kondisi tersebut tercermin pada data penelitian ini: dari 8.000 ulasan gabungan yang dihimpun (masing-masing 4.000 dari Brainly dan Ruang Guru) untuk rentang Mei-Juni 2026, setelah pembersihan duplikasi jumlahnya menyusut menjadi

This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

6.151 ulasan, terbagi atas 2.836 ulasan Brainly dan 3.315 ulasan Ruang Guru. Memeriksa ribuan teks tersebut satu demi satu jelas tidak praktis, sehingga diperlukan analisis sentimen otomatis berbasis pembelajaran mesin untuk memotret opini pengguna atas kualitas layanan kedua aplikasi. Kesulitan lain muncul dari ragam bahasa Indonesia informal pada ulasan, yang dipenuhi singkatan, kesalahan ketik, dan campur kode, sehingga pembersihan teks konvensional kurang memadai. Selain itu, label otomatis berdasarkan rating bintang memunculkan ketidakseimbangan kelas, yakni 79,8% positif berbanding 20,2% negatif, kondisi yang umumnya membuat model klasifikasi condong ke kelas mayoritas. Representasi fitur TF-IDF dari korpus ini pula menghasilkan ruang fitur berdimensi tinggi sekaligus jarang terisi, dengan sparsitas hingga 99,78%, yang menuntut kecermatan dalam pemilihan *hiperparameter*[1].

Beberapa kajian terdahulu sudah menerapkan *Naive Bayes*, *K-Nearest Neighbor*, dan *Decision Tree* untuk analisis sentimen ulasan aplikasi sejenis, kendati belum ditemukan yang memperbandingkan ketiganya secara khusus pada ranah aplikasi edukasi di Indonesia. Riset terhadap ulasan aplikasi DANA mencatat *Naive Bayes* memperoleh akurasi 80,80% dalam memilah sentimen positif dan negatif [1]. Pada ulasan aplikasi Wondr, metode sejenis menghasilkan akurasi 82%, meski performanya pada kelas minoritas tetap terbatas [2]. Kajian lain terhadap ulasan aplikasi Identitas Kependudukan Digital memperlihatkan algoritma KNN pada k=17 mencapai akurasi 82%, presisi 79%, dan recall 82%, namun masih sulit mengenali kelas minoritas [3]. Di ranah aplikasi investasi, gabungan *Naive Bayes* dan KNN dilaporkan cukup efektif mengklasifikasikan ulasan aplikasi Bibit, walau riset tersebut belum melibatkan *Decision Tree* sebagai pembanding [4]. Studi terhadap ulasan BCA Mobile yang memakai *Decision Tree* memperoleh akurasi 77,11% dengan rincian sentimen 72,0% positif, 21,6% negatif, dan 6,4% netral, dan turut menemukan penurunan performa pada kelas minoritas akibat ketidakseimbangan data [5]. Riset yang membandingkan ketiga algoritma pada ulasan Threads dan Twitter menampilkan hasil bervariasi menurut platform, mengindikasikan kinerja setiap algoritma dipengaruhi karakteristik korpus yang dipakai [6]. Perbandingan *Naive Bayes* dengan Support Vector Machine pada aplikasi LinkedIn menunjukkan akurasi 87,46% berbanding 88,48%, membuktikan *Naive Bayes* tetap mampu bersaing dengan metode yang lebih rumit [7]. Riset terhadap ulasan pembelajaran daring turut melaporkan ketiga algoritma dapat diterapkan bersama untuk mengelompokkan sentimen [8], sedangkan pada ulasan aplikasi Gojek, perbandingan ketiganya konsisten menempatkan *Naive Bayes* sebagai algoritma paling unggul (89%), disusul KNN (86%) dan *Decision Tree* (84%) [9].

Bertolak dari rangkaian riset tersebut, terlihat celah yang belum banyak terjawab: kebanyakan riset hanya memakai satu atau dua algoritma klasifikasi sehingga belum tersedia gambaran menyeluruh mengenai kinerja relatif ketiganya pada *dataset* yang identik, dan hampir tidak ada yang secara tegas mengukur dampak penyeimbangan data lewat SMOTE [10] terhadap akurasi akhir model. Padahal ketidakseimbangan label hampir selalu muncul ketika label sentimen diturunkan dari rating bintang, sebagaimana tampak pula pada dataset Brainly dan Ruang Guru dalam riset ini. Kebaruan riset ini terletak pada penggabungan tiga unsur sekaligus: perbandingan langsung *Naive Bayes*, KNN, dan *Decision Tree* pada ranah aplikasi edukasi Indonesia yang belum banyak dikaji secara komparatif; pengujian empiris pengaruh SMOTE terhadap performa ketiga algoritma pada data TF-IDF berdimensi tinggi; serta optimasi *hiperparameter* sistematis lewat *GridSearchCV* dan *StratifiedKFold* yang dilengkapi telaah kata kunci pembeda sentimen khusus ranah edukasi.

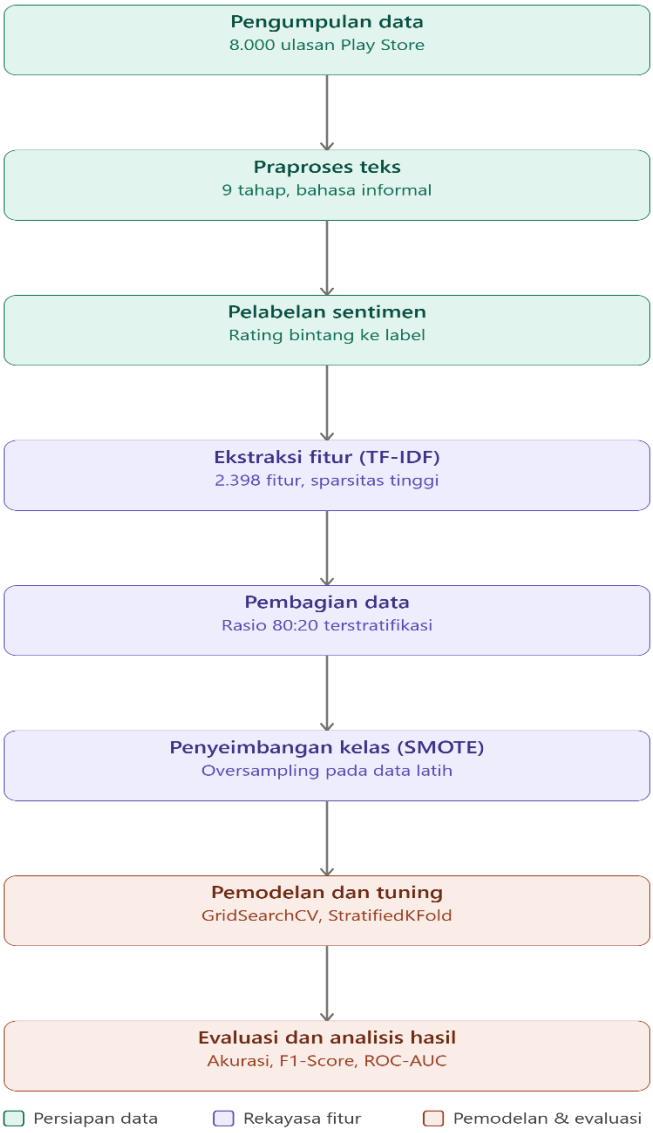
Mengacu pada paparan tersebut, riset ini disusun untuk: (1) memperbandingkan performa *Naive Bayes*, KNN, dan *Decision Tree* dalam mengklasifikasikan sentimen ulasan pengguna Brainly dan Ruang Guru; (2) menelaah pengaruh SMOTE serta *optimasi hiperparameter* terhadap akurasi ketiga model; dan (3) memetakan perbedaan distribusi sentimen beserta kata kunci dominan antara kedua aplikasi sebagai bahan evaluasi layanan. Temuan riset ini diharapkan memberi rekomendasi metodologis bagi peneliti sejenis sekaligus masukan praktis bagi pengembang kedua aplikasi.

2. METODOLOGI PENELITIAN

Penelitian ini dilakukan dalam delapan tahap yang berurutan, yaitu pengumpulan data , pembersihan dan *pra-pemrosesan teks*, *pelabelan sentimen* , *ekstraksi fitur*, pembagian data latih dan uji, proses *sinkronisasi* kelas, pencetakan dan pengenal *parameter hiper*, serta *evaluasi* dan analisis hasil.Tahap penelitian yakini:

Copyright © 2023 **Author**, Page 1112
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294



Gambar 1. Tahap metode penelitian

2.1 Pengumpulan Data

Data dihimpun memakai pustaka *google-play-scraper* dengan mode pengambilan bertahap (200 ulasan per permintaan), *parameter* bahasa Indonesia, dan urutan ulasan terbaru. Dua aplikasi yang dijadikan objek adalah Brainly dan Ruang Guru, masing-masing ditarget 4.000 ulasan untuk periode Mei-Juni 2026, sehingga total data mentah mencapai 8.000 baris sebelum pembersihan duplikasi.

2.2 Praproses Teks

Mengingat gaya bahasa ulasan yang *informal*, naskah diolah lewat sembilan langkah berurutan: penyamaan huruf kecil, penghapusan tautan, penghapusan emoji, penghapusan angka, penghapusan tanda baca, normalisasi kata tidak baku memakai kamus mandiri 163 entri, tokenisasi, penyaringan kata henti gabungan *Sastraw*i-NLTK sejumlah 861 kata, dan penyederhanaan kata berimbuhan lewat algoritma *stemming Sastrawi* [3], [5]. Pendekatan berjenjang ini lazim dipakai pada riset ulasan aplikasi berbahasa Indonesia agar noise tekstual berkurang tanpa menghapus makna sentimen.

2.3 Pelabelan Sentimen

Karena *Play Store* tidak menyertakan label sentimen secara langsung, label diturunkan dari rating bintang yang diberikan pengguna: rating 1 sampai 3 dikategorikan negatif, sedangkan rating 4 dan 5 dikategorikan positif, mengikuti kelaziman pada riset sejenis [9]. Aturan ini menempatkan rating tengah (3) sebagai sinyal ketidakpuasan parsial, bukan diabaikan begitu saja.

Copyright © 2023 **Author**, Page 1113
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

2.4 Ekstraksi Fitur (TF-IDF)

Teks bersih diubah ke representasi numerik lewat *Term Frequency-Inverse Document Frequency* dengan batas 5.000 fitur, kombinasi *unigram-bigram*, ambang *frekuensi* dokumen minimum tiga dan maksimum 90 persen, serta pembobotan sublinear agar kata yang sangat sering muncul tidak mendominasi bobot fitur.

2.5 Pembagian Data

Matriks fitur dan label dibagi dengan rasio 80 banding 20 secara terstratifikasi dan kunci acak tetap, sehingga proporsi kelas positif dan negatif tetap konsisten pada data latih maupun data uji, sekaligus mencegah data uji ikut "terlihat" oleh model selama pelatihan maupun saat penyeimbangan kelas dilakukan.

2.6 Penyeimbangan Kelas (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) diterapkan khusus pada data latih untuk mengatasi dominasi kelas positif, lewat interpolasi sintetis antartetangga terdekat kelas minoritas pada ruang fitur, sehingga jumlah sampel kedua kelas pada data latih menjadi setara tanpa mengubah komposisi data uji yang tetap merepresentasikan kondisi asli di lapangan.

2.7 Pemodelan dan Tuning

Tiga algoritma yang diperbandingkan adalah *Multinomial Naive Bayes*, *K-Nearest Neighbor*, dan *Decision Tree*, masing-masing diuji dalam tiga kondisi: model dasar tanpa penyeimbangan, model dengan SMOTE, dan model hasil pencarian *hiperparameter* terbaik memakai *GridSearchCV* yang dipadukan validasi silang *StratifiedKFold* lima lipatan agar pemilihan parameter tidak bias pada satu pembagian data saja dan tetap menjaga proporsi kelas di setiap lipatan.

2.8 Evaluasi dan analisis Hasil

Kinerja akhir tiap model diukur lewat *Accuracy*, *Precision*, *Recall*, *F1-Score* tertimbang, *Matthews Correlation Coefficient*, dan ROC-AUC, dilengkapi validasi silang untuk memeriksa kestabilan model di luar satu kali pembagian data uji. Hasil dari kedelapan tahap ini selanjutnya dibahas lebih rinci pada bagian Hasil dan Pembahasan.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan dan Praproses Data

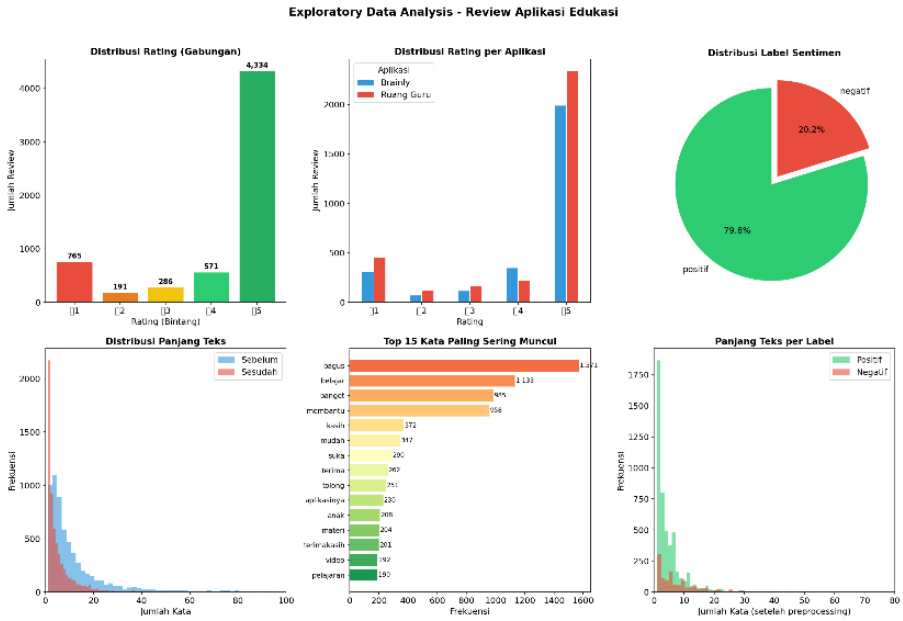
Sebelum membahas hasil klasifikasi, bagian ini terlebih dahulu memaparkan kualitas data mentah dan dampak praproses terhadapnya, karena keduanya menentukan keandalan analisis pada tahap selanjutnya. Proses scraping menghimpun 8.000 ulasan mentah yang setelah pembersihan duplikasi menyusut menjadi 6.151 ulasan bersih. Pemeriksaan nilai kosong pada seluruh kolom data tidak menemukan satu pun nilai hilang, menandakan kualitas hasil scraping yang baik. Contoh transformasi praproses memperlihatkan bagaimana pipeline sembilan tahap menyederhanakan teks tanpa kehilangan makna sentimen: ulasan "sangat membantu bagi pelajar atau pekerja, btw terus semangat ya jalani hidup nya" tereduksi menjadi "membantu pelajar pekerja btw semangat jalani hidup nya", sementara ulasan "Jelek! Banyak bug, sering error, aplikasi gak berguna" tersaring bersih menjadi "jelek bug error berguna". Secara agregat, rata-rata panjang ulasan menyusut dari 11,79 kata menjadi 5,86 kata setelah praproses, dengan median turun dari 7 kata menjadi 3 kata, sebuah indikasi bahwa sebagian besar kata yang terbuang adalah kata fungsi yang memang tidak membawa muatan sentimen.

3.1 Destribusi dan Pelabelan Sentimen

Penerapan aturan konversi rating-ke-label terhadap 6.151 ulasan menghasilkan 4.911 ulasan berlabel positif (79,84%) dan 1.240 ulasan berlabel negatif (20,16%). Gambar 2 merangkum enam aspek distribusi data sekaligus: pola rating bintang yang condong ke angka 4 dan 5, proporsi rating per aplikasi, pembagian label sentimen, perbandingan panjang teks sebelum dan sesudah praproses, lima belas kata yang paling sering muncul, serta sebaran panjang teks menurut label.

Copyright © 2023 **Author**, Page 1114
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

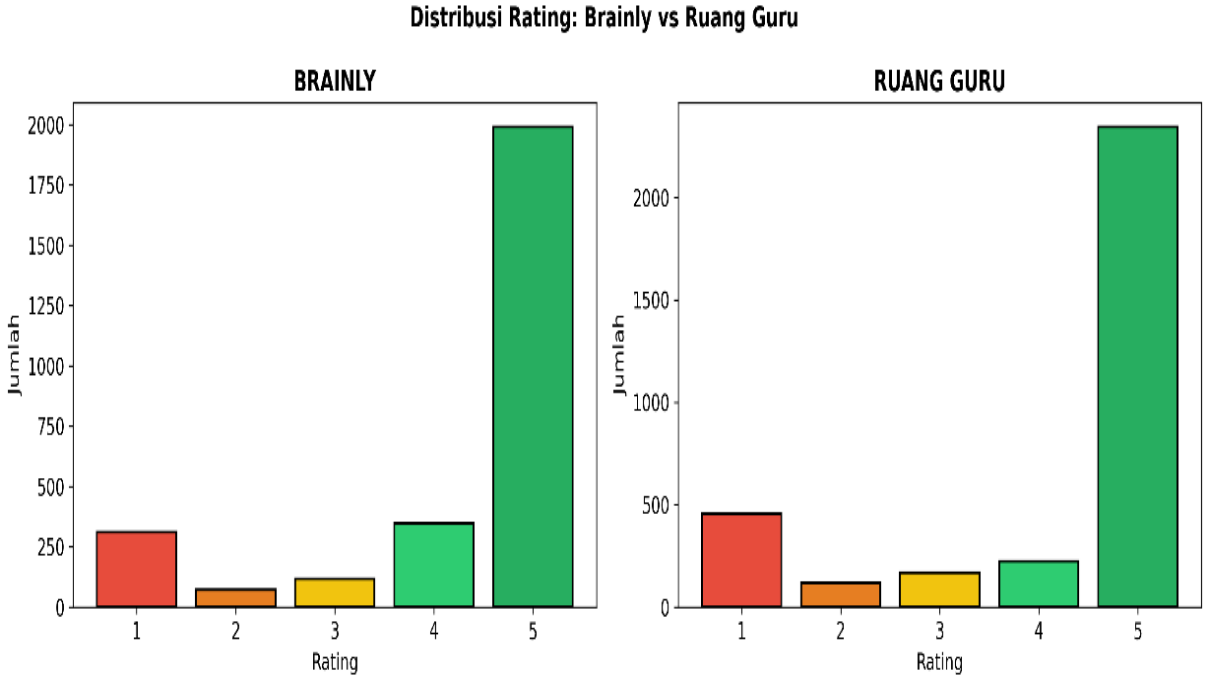


Gambar 2. Eda Destribusi

Pola pada panel distribusi panjang teks memperkuat temuan di subbab sebelumnya, sementara panel kata tersering menunjukkan bahwa kosakata seperti "bagus" dan "membantu" muncul jauh lebih dominan dibanding kosakata bernuansa negatif, sejalan dengan proporsi kelas yang memang condong positif.

3.2 Perbandingan Karakteristik Brainly dan Ruang Guru

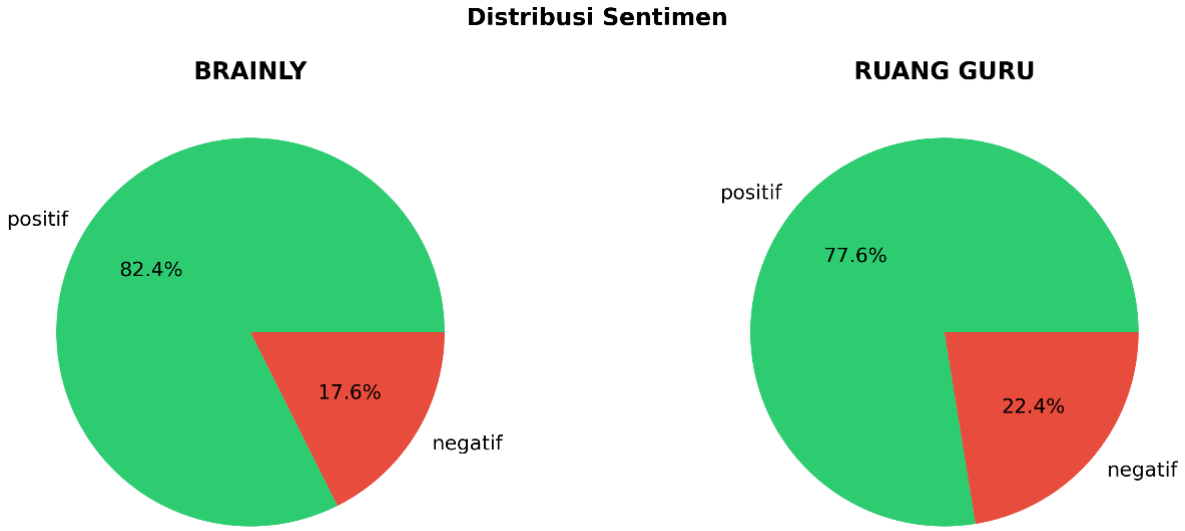
Gambar 3 menampilkan distribusi rating bintang untuk kedua aplikasi. Rata-rata rating Brainly (4,28 dari 5) sedikit lebih tinggi dibanding Ruang Guru (4,18 dari 5).



Gambar 3. *Destribusi Rating*

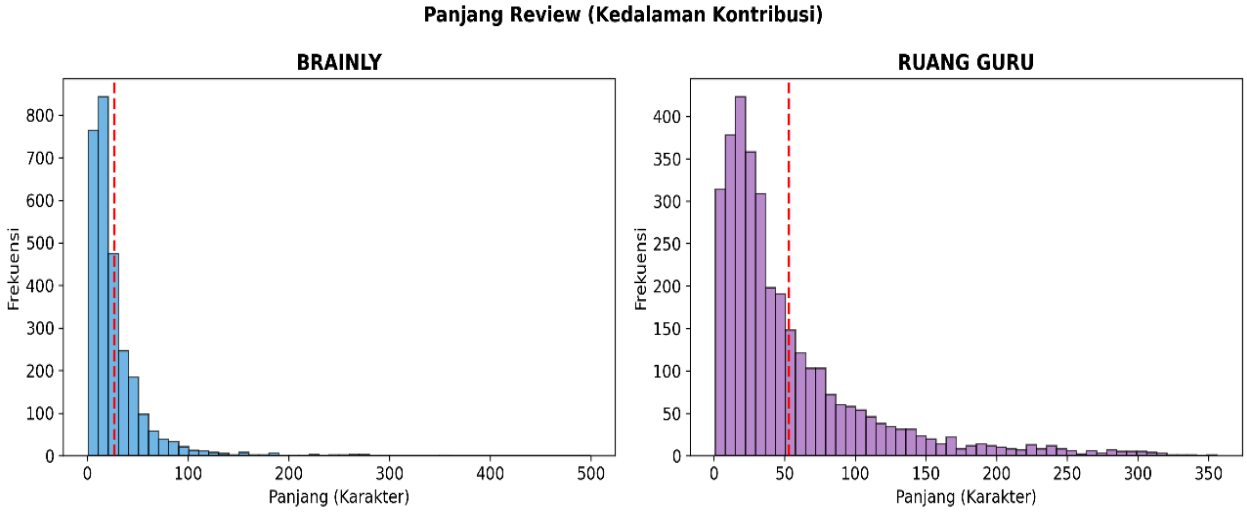
Copyright © 2023 **Author**, Page 1115
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294



Gambar 4. *Destribusi Sentimen*

Dipecah per aplikasi, Brainly mencatat 2.338 ulasan positif dan 498 ulasan negatif (82,4% berbanding 17,6%), sementara Ruang Guru mencatat 2.573 ulasan positif dan 742 ulasan negatif (77,6% berbanding 22,4%). Proporsi sentimen negatif Ruang Guru lebih tinggi 4,8 poin *persentase* dibanding Brainly. Temuan ini sejalan dengan karakter model bisnis Ruang Guru yang menerapkan skema freemium, di mana keterbatasan akses fitur gratis kerap memicu ketidakpuasan, sementara Brainly yang berbasis forum tanya-jawab komunitas relatif lebih jarang memicu keluhan terkait biaya.



Gambar 5. Panjang *Review*

Gambar 5 menampilkan distribusi panjang ulasan dalam karakter untuk kedua aplikasi. Gambar ini memperlihatkan rata-rata ulasan Ruang Guru (53 karakter) hampir dua kali lebih panjang dibanding ulasan Brainly (27 karakter), mengindikasikan pengguna Ruang Guru cenderung menulis ulasan yang lebih elaboratif. Perbedaan ini masuk akal mengingat audiens Ruang Guru kerap melibatkan orang tua yang menilai pengalaman belajar anak secara berkelanjutan, sementara pengguna Brainly umumnya pelajar yang menulis ulasan singkat segera setelah mendapat bantuan jawaban.

3.3 Pola Kata Kunci Pada Data Mentah

Gambar 6 menyajikan wordcloud sentimen positif dan negatif, sementara Gambar 7 merinci dua puluh kata dengan frekuensi tertinggi pada masing-masing sentimen dalam bentuk diagram batang.

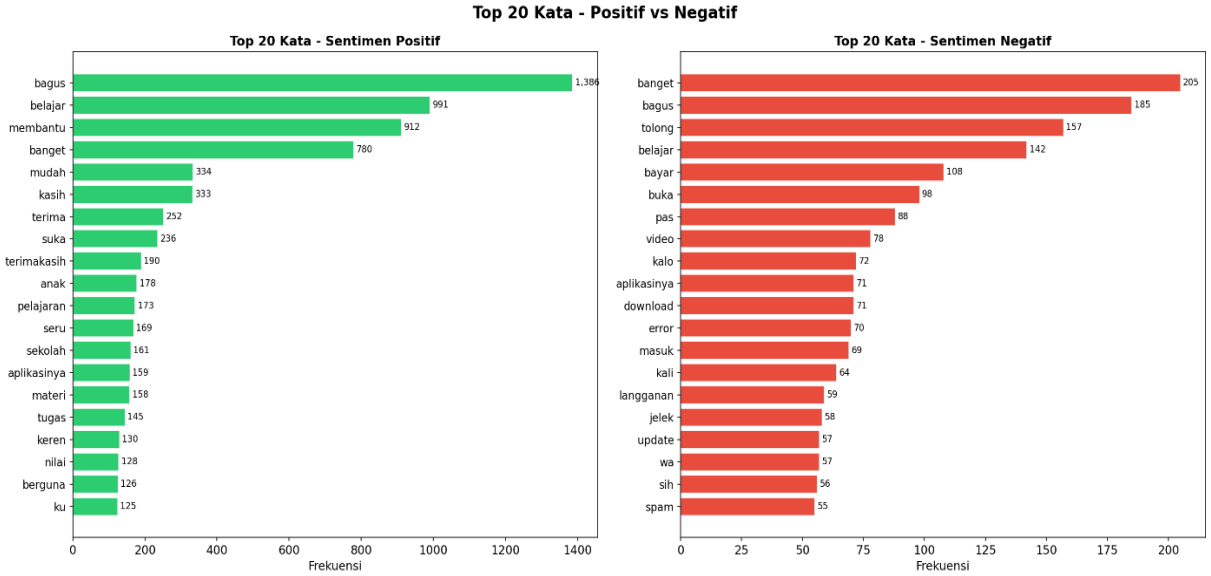
Copyright © 2023 **Author**, Page 1116
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294



Gambar 6. *Worldcloud Sentimen*

Gambar 7 merinci dua puluh kata dengan frekuensi tertinggi pada masing-masing sentimen dalam bentuk diagram batang. Pada sentimen positif, kata "bagus", "membantu", "banget", dan "mudah" mendominasi baik di Brainly maupun Ruang Guru, namun kata "jawabannya" dan "tugas" lebih menonjol di Brainly sebagai cermin interaksi tanya-jawab seketika, sedangkan "seru", "suka", dan "anak" lebih menonjol di Ruang Guru sebagai cermin pembelajaran terstruktur yang melibatkan orang tua.



Gambar 7. *Top* Kata *Sentimen*

Pada sentimen negatif, kata-kata seperti "iklan", "error", dan "lambat" muncul konsisten di kedua aplikasi, mengindikasikan keluhan teknis sebagai sumber ketidakpuasan yang umum di seluruh kategori aplikasi edukasi. Pola ini menunjukkan bahwa meski model bisnis kedua aplikasi berbeda, ekspektasi dasar pengguna terhadap kestabilan teknis tetap sama, dan kegagalan memenuhi ekspektasi tersebut memunculkan keluhan yang serupa di kedua platform.

3.4 Ekstraksi Fitur dan Pembagian Data

Penerapan TF-IDF terhadap 6.151 dokumen menghasilkan matriks berukuran 6.151 × 2.398, lebih kecil dari batas maksimum 5.000 fitur karena kosakata unik dalam korpus yang telah melalui praproses tidak mencapai jumlah tersebut. Tingkat sparsitas matriks ini mencapai 99,78%, karakteristik umum pada data teks berdimensi tinggi yang berarti hanya sekitar 0,22% sel pada matriks berisi nilai bukan nol. Pembagian data 80:20 secara terstratifikasi menghasilkan 4.920 data

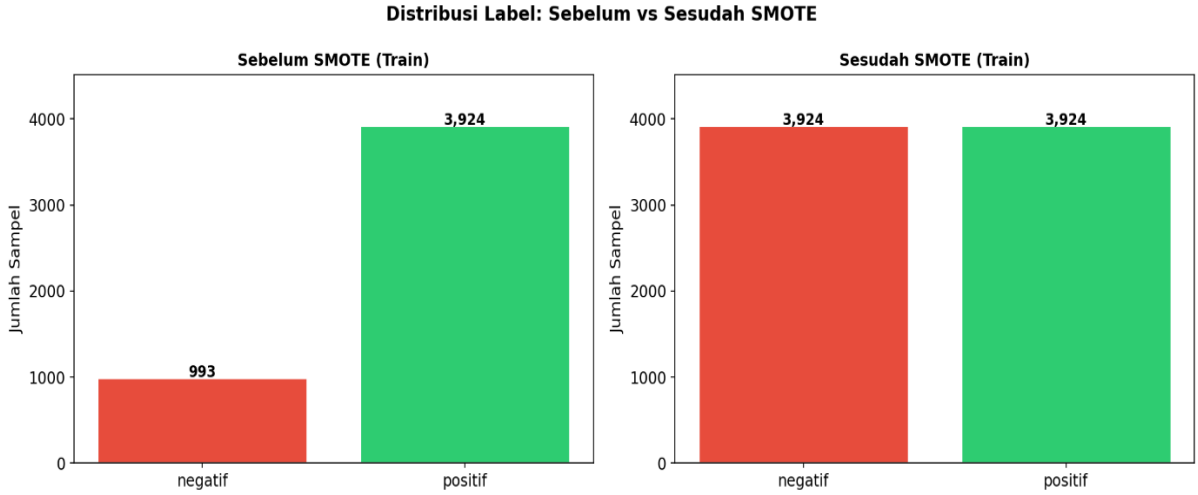
Copyright © 2023 **Author**, Page 1117
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

latih (992 negatif dan 3.928 positif) serta 1.231 data uji (248 negatif dan 983 positif), dengan proporsi kelas yang konsisten terjaga pada kedua subset.

3.6 Penanganan Ketidakseimbangan Kelas dengan SMOTE

Penerapan SMOTE pada data latih mengubah distribusi kelas dari 992 berbanding 3.928 menjadi seimbang sempurna 3.928 berbanding 3.928, menghasilkan total 7.856 sampel data latih, meningkat 2.936 sampel sintetis pada kelas negatif.



Gambar 8. SMOTE Destribusi

Gambar 8 memperlihatkan perubahan ini secara berdampingan, di mana batang kelas negatif pada panel kanan tampak setara dengan batang kelas positif. Penting dicatat bahwa proses ini hanya diterapkan pada data latih agar data uji tetap merepresentasikan distribusi asli dunia nyata, sehingga hasil evaluasi tidak bias secara artifisial.

3.7 Hasil Klasifikasi pada Sembilan Skenario Pengujian

Table 1 merangkum performa ketiga algoritma pada sembilan skenario: model dasar tanpa SMOTE, model dasar dengan SMOTE, dan model hasil tuning dengan SMOTE.

Table 1. Ringkasan Evaluasi Sembilan Skenario Pengujian

Model	Accuracy	Precision	Recall	F1-Score	MCC	ROC-AUC
NB (Dasar, tanpa SMOTE)	88,95%	88,57%	88,95%	88,05%	0,6256	90.99%
KNN (Dasar, tanpa SMOTE)	81,40%	78,68%	81,40%	76,92%	0,2650	67,85%
DT (Dasar, tanpa SMOTE)	83,59%	83,03%	83,59%	83,27%	0,4718	76,28%
NB (Dasar, SMOTE)	82,78%	87,53%	82.78%	84,02%	0,5852	91,32%
KNN (Dasar, SMOTE)	52,43%	82,46%	57,43%	61,04%	0,3242	78,42%
DT (Dasar, SMOTE)	82,78%	82,88%	82,78%	82,83%	0,4680	77,13%
NB (Tuning, SMOTE)	82,78%	86,05%	82,78%	83,79%	0,5517	88,44%
KNN (Tuning, SMOTE)	77,66%	81,82%	77,66%	79,07%	0,4214	78,51%
DT (Tuning, SMOTE)	82,21%	82,74%	82,21%	82,45%	0,4630	77,44%

Optimasi hiperparameter menghasilkan konfigurasi terbaik: Naive Bayes dengan *alpha*=0,01, KNN dengan metrik *cosine*, *n_neighbors*=3, dan *weights=distance*, serta Decision Tree dengan kriteria *entropy*, *max_depth=None*, dan *min_samples_split*=5. Menariknya, skor validasi silang lima lipatan pada ketiga model tuning hasil SMOTE berada pada rentang yang berdekatan, yakni 86,29% hingga 86,60%, jauh lebih homogen dibanding rentang skor pada data uji tunggal. Hal ini mengindikasikan bahwa stabilitas model selama validasi silang tidak otomatis berarti performa yang setara saat diuji pada data uji yang tidak pernah dilihat sebelumnya, sebuah pengingat bahwa kedua jenis pengujian sebaiknya selalu dilaporkan bersamaan, bukan saling menggantikan.

Temuan paling menarik dari tabel ini justru muncul dari pola yang tidak sejalan dengan asumsi umum bahwa SMOTE selalu memperbaiki performa pada data tidak seimbang. Pada Naive Bayes, penerapan SMOTE menurunkan akurasi sebesar 6,17 poin, walau F1-Score sedikit lebih tinggi karena keseimbangan presisi-*recall* antarkelas membaik. Decision Tree mengalami penurunan akurasi yang lebih ringan, sementara KNN mengalami kemerosotan paling drastis, akurasinya terjun dari 81,40% menjadi hanya 57,43% ketika SMOTE diterapkan tanpa optimasi lanjutan, sebelum terangkat kembali ke 77,66% setelah *tuning*, meski tetap di bawah model dasar tanpa SMOTE. Pola ini dapat dijelaskan dari karakteristik data:

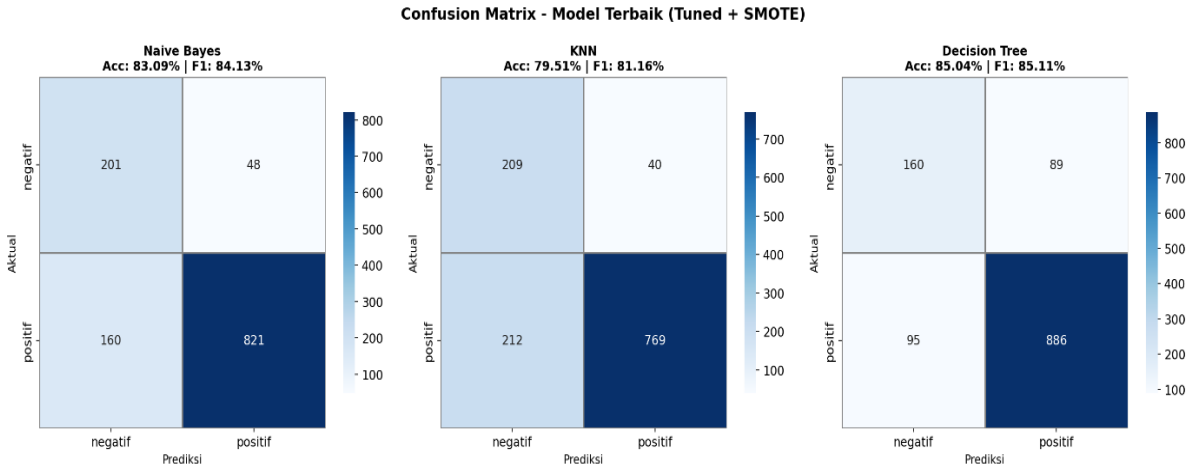
Copyright © 2023 **Author**, Page 1118
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

matriks TF-IDF berdimensi tinggi dengan sparsitas ekstrem membuat interpolasi sintetis SMOTE antara dua sampel minoritas yang berjarak jauh cenderung menghasilkan titik data yang berada di area tumpang tindih antarkelas, alih-alih memperkuat batas keputusan kelas minoritas secara representatif. KNN paling rentan terhadap efek ini karena keputusannya murni berdasarkan jarak ke tetangga terdekat, sehingga satu sampel sintetis yang "mengambang" di antara dua kelas langsung mengacaukan voting mayoritas, berbeda dengan Naive Bayes yang mengandalkan estimasi probabilitas independen per kelas dan relatif lebih tahan terhadap distorsi lokal semacam ini [10].

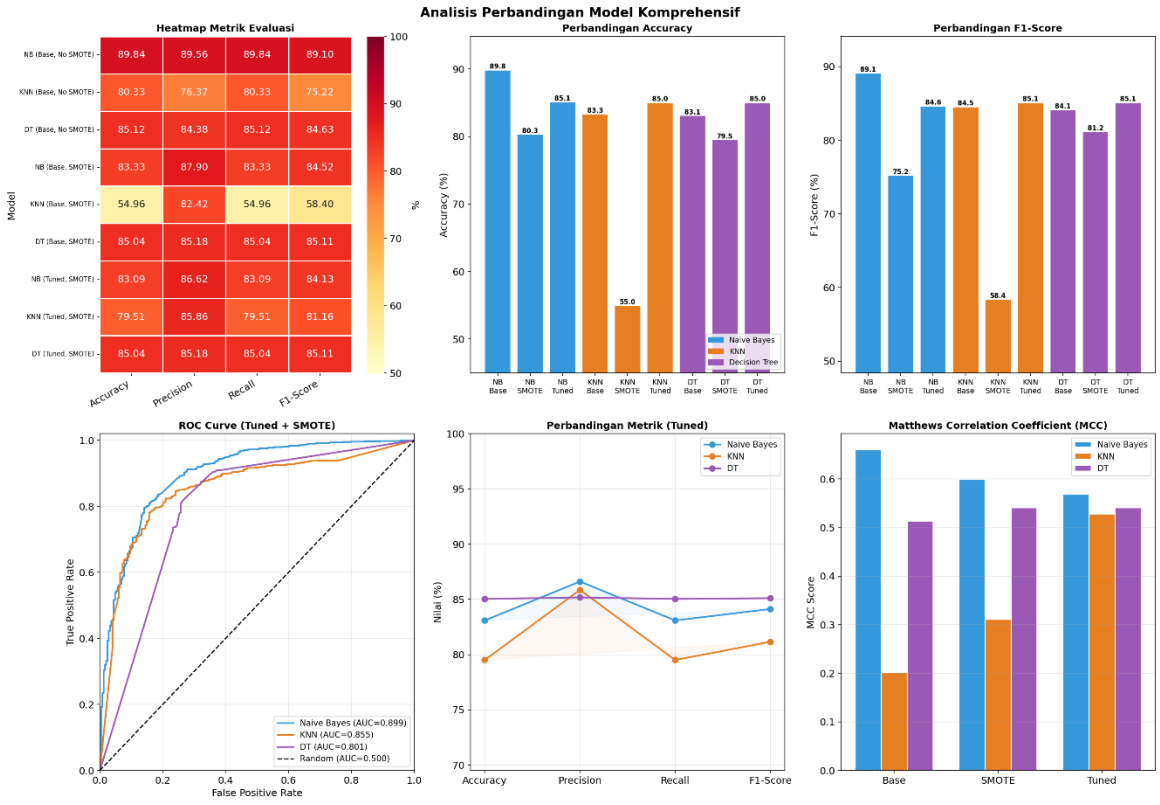
3.8 Evaluasi Visual: Confusion Matrix dan Perbandingan Model

Gambar 9 menampilkan confusion matrix ketiga model pada skema tuning dengan SMOTE, memperlihatkan bahwa kesalahan klasifikasi paling banyak terjadi pada kelas negatif yang diprediksi sebagai positif, konsisten dengan sifat data yang tetap condong ke kelas mayoritas pada da ta uji.



Gambar 9. Confusion Matrik

Gambar 10 memadukan enam panel sekaligus: heatmap metrik evaluasi seluruh skenario, perbandingan akurasi dan F1-Score antarmodel, kurva ROC, grafik radar metrik, serta perbandingan skor Matthews Correlation Coefficient.



Gambar 10. perbandingan Model

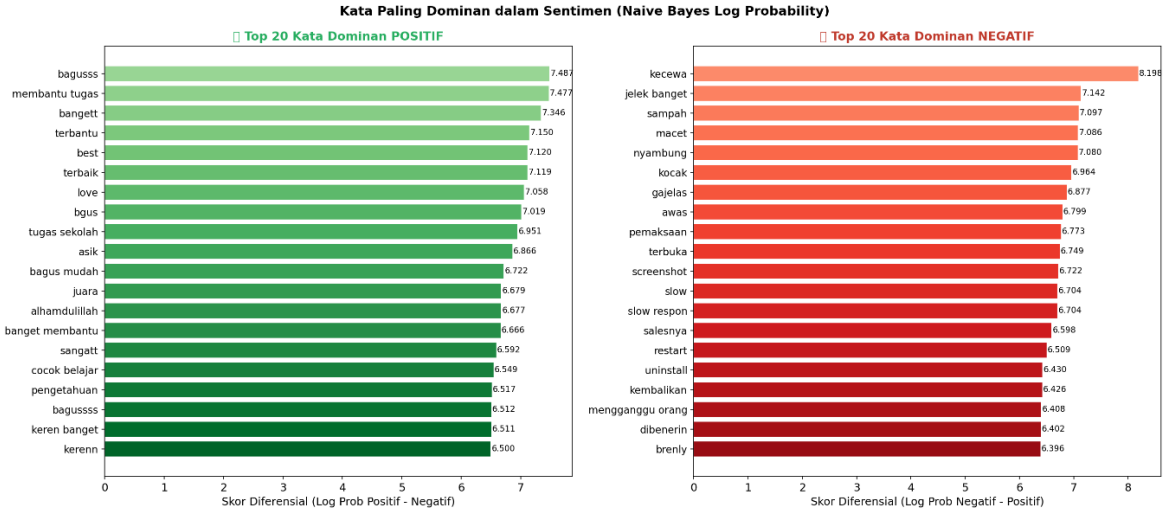
Copyright © 2023 **Author**, Page 1119
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

Pada kurva ROC, Naive Bayes mencatat AUC tertinggi, menegaskan kemampuannya memisahkan kelas positif dan negatif lebih baik dibanding KNN dan Decision Tree meski selisih akurasinya dengan Decision Tree tidak besar. Grafik radar pada panel kelima memperlihatkan Naive Bayes unggul hampir di seluruh sumbu metrik, sementara KNN secara konsisten berada pada sumbu terluar yang menunjukkan performa terendah.

3.9 Kata Kunci Pembeda Sentimen Menurut Model

Gambar 11 menyajikan dua puluh kata dengan skor diferensial log probability tertinggi pada model Naive Bayes, dipisah untuk sentimen positif dan negatif. Kata-kata seperti "best", "membantu tugas", "terbaik", "terbantu", dan "mudah pahami" memiliki skor diferensial tertinggi pada sentimen positif, konsisten merujuk pada manfaat akademik langsung yang dirasakan pengguna. Pada sisi negatif, kata-kata dominan meliputi "sampah", "gajelas", "bayar mahal", "macet", dan "pemaksaan", mengindikasikan tiga sumber keluhan utama: kualitas konten yang dianggap buruk, masalah teknis aplikasi, dan keberatan terhadap model berbayar.



Gambar 11. feature importance

Berbeda dengan Gambar 7 yang menghitung frekuensi kata mentah, Gambar 11 mengukur seberapa kuat sebuah kata membedakan kedua kelas menurut model, sehingga kata yang sering muncul namun netral makna sentimennya tidak ikut menonjol di sini. Perbedaan dua sudut pandang ini saling melengkapi: frekuensi mentah berguna untuk memahami topik apa yang paling banyak dibicarakan pengguna, sementara skor diferensial lebih berguna untuk memahami kata mana yang paling bisa diandalkan sebagai sinyal sentimen oleh model klasifikasi.

3.10 Analisis Kesalahan Klasifikasi

Pengujian model Naive Bayes ber-tuning pada 1.231 data uji menghasilkan 212 prediksi yang keliru, atau tingkat kesalahan sebesar 17,22%. Penelusuran terhadap kasus kesalahan ini mengungkap pola yang berulang: mayoritas kesalahan terjadi pada ulasan berlabel positif namun diprediksi negatif, biasanya karena teks setelah praproses menjadi sangat singkat dan kehilangan konteks emosional, misalnya ulasan "SERUUUU" yang setelah pembersihan tersisa tanpa kata sentimen baku yang dikenali model, atau ulasan "mayan" yang terlalu pendek untuk memberi sinyal probabilitas kuat. Kasus lain melibatkan ulasan yang memuat keluhan teknis namun tetap diberi rating tinggi karena penilaian keseluruhan pengguna tetap positif, misalnya ulasan yang menyebut "tolong perbaikan masalah keluar tiba tiba atau aplikasi error" namun tetap diberi bintang lima, situasi yang menunjukkan keterbatasan inheren pendekatan pelabelan berbasis rating semata, di mana teks dan skor numerik tidak selalu konsisten satu sama lain. Pola kesalahan semacam ini sulit diperbaiki hanya lewat penyesuaian algoritma, karena akar masalahnya terletak pada proses pelabelan itu sendiri, bukan pada kemampuan model mempelajari pola dari data yang tersedia.

3.11 Perbandingan dengan penelitian Terdahulu

Table 2 memosisikan temuan penelitian ini terhadap penelitian sejenis pada ulasan aplikasi lain.

Table 2. Perbandingan Akurasi dengan Penelitian Terdahulu

Penelitian	Objek	Algoritma	Akurasi terbaik
[1]	Aplikasi Dana	Naïve Bayes	80,80%
[2]	Aplikasi Wondr	Naïve Bayes	82%
[3]	Aplikasi IKD	KNN	82%

Copyright © 2023 **Author**, Page 1120
This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

[5]	BCA Mobile	Decision Tree	77,11%
[9]	Aplikasi Gojek	NB/KNN/DT	89%(NB)
Penelitian ini	Brainly dan Ruang Guru	NB/KNN/DT	88,95%(NB dasar)

Akurasi Naive Bayes dasar pada penelitian ini (88,95%) berada pada rentang atas dibanding studi sejenis, mendekati hasil pada aplikasi Gojek [9] dan melampaui hasil pada aplikasi DANA [1], Wondr [2], maupun Identitas Kependudukan Digital [3]. Konsistensi keunggulan Naive Bayes dibanding KNN dan Decision Tree pada penelitian ini juga sejalan dengan temuan pada aplikasi Gojek [9], memperkuat indikasi bahwa algoritma ini secara umum lebih sesuai untuk data ulasan aplikasi.

3.12 Implikasi Praktis

Bagi pengembang Brainly, temuan kata kunci negatif yang berkutat pada kualitas jawaban dan keandalan teknis mengindikasikan perlunya mekanisme moderasi atau verifikasi jawaban yang lebih ketat. Bagi Ruang Guru, proporsi sentimen negatif yang lebih tinggi serta kemunculan kata kunci terkait harga menunjukkan perlunya evaluasi ulang terhadap struktur freemium dan transparansi fitur berbayar. Bagi peneliti yang berencana melakukan studi sejenis, temuan mengenai efek SMOTE yang tidak selalu positif pada data TF-IDF berdimensi tinggi menjadi catatan metodologis penting agar pengujian model dasar tetap disertakan sebagai pembanding, bukan dilewati dengan asumsi teknik penyeimbangan data pasti meningkatkan performa. Secara umum, kombinasi tabel evaluasi kuantitatif dan analisis kata kunci kualitatif yang digunakan dalam penelitian ini dapat dijadikan kerangka kerja yang relatif mudah direplikasi pada domain ulasan aplikasi lain di luar sektor edukasi.

4. KESIMPULAN

Penelitian ini membandingkan performa tiga algoritma klasifikasi Naïve Bayes, K-Nearest Neighbor, dan Decision Tree dalam analisis sentimen 6.151 ulasan aplikasi Brainly dan Ruang Guru dari Google Play Store. Hasil menunjukkan Naïve Bayes mencapai performa terbaik dengan akurasi 88,95% tanpa optimasi SMOTE, sementara dengan optimasi mencapai akurasi 82,78%, F1-*Score* 83,79%, dan ROC-AUC 88,44%. Temuan ini membuktikan bahwa penyeimbangan data tidak otomatis meningkatkan performa dan perlu pengujian empiris terhadap model dasar. Dari sisi distribusi sentimen, Ruang Guru menunjukkan ulasan negatif lebih tinggi (22,4%) dibanding Brainly (17,6%), sejalan dengan model bisnis mereka. Ruang Guru berbayar menghadapi keluhan harga, sedangkan Brainly berbasis komunitas menghadapi keluhan kualitas jawaban. Analisis kata kunci menunjukkan perbedaan jelas antara ulasan positif (membantu, mudah, terbaik) dan negatif (sampah, mahal, *error*). Penelitian memiliki keterbatasan: pelabelan sentimen bergantung pada rating bintang saja dan cakupan data hanya dua bulan. Penelitian lanjutan disarankan menggunakan model deep learning seperti IndoBERT dan analisis sentimen berbasis aspek untuk hasil lebih granular.

REFERENCES

[1] R. F. Chandra and D. A. Putri, "Implementasi Metode Naïve Bayes Pada Ulasan Pengguna Aplikasi Dana Di Google Play Store," Jurnal Infortech, vol. 7, no. 1, pp. 64–69, 2025, doi: 10.31294/infortech.v7i1.12366.

[2] S. Nurhikmah, R. Ramadani, and G. Triyono, "Analisis Sentimen pada Ulasan Aplikasi Wondr di Play Store dengan Metode Naive Bayes," Jurnal Algoritma, vol. 22, no. 2, pp. 1919–1930, 2025, doi: 10.33364/algoritma/v.22-2.2507.

[3] M. Ulfa, R. H. Kusumodestoni, and A. Sucipto, "Analisis Sentimen Review Aplikasi Identitas Kependudukan Digital di Google Play Store Menggunakan KNN," Jurnal Informatika Teknologi dan Sains (Jinteks), vol. 6, no. 4, pp. 1155–1165, 2024, doi: 10.51401/jinteks.v6i4.4963.

[4] A. Azmi, Y. Hendriyani, I. P. Dewi, dan K. Budayawan, "Analisis Sentimen Ulasan Pengguna Aplikasi Bibit Menggunakan Algoritma Naive Bayes dan K-Nearest Neighbors (KNN)," Jurnal Pendidikan Tambusai, vol. 9, no. 2, pp. 14040–14048, 2025, doi: 10.31004/jptam.v9i2.27464.

[5] A. D. Sugiarto and M. S. Utomo, "Analisis Sentimen Ulasan Pengguna BCA Mobile di Google Play Store Menggunakan Metode Decision Tree," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 5, 2025, doi: 10.36040/jati.v9i5.14969.

[6] M. Iqbal, A. D. Wiranata, R. Suwito, and R. F. Ananda, "Perbandingan Algoritma Naïve Bayes, KNN, dan Decision Tree terhadap Ulasan Aplikasi Threads dan Twitter," KLIK: Kajian Ilmiah Informatika dan Komputer, vol. 4, no. 3, pp. 1799–1807, 2023, doi: 10.30865/klik.v4i3.1402.

[7] G. S. Al-Husna, D. Asmarajati, I. A. Ihsannuddin, and R. Mahmudati, "Perbandingan Metode Naïve Bayes dan Support Vector Machine untuk Analisis Sentimen pada Ulasan Pengguna Aplikasi LinkedIn," STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer, vol. 3, no. 2, pp. 139–144, 2024, doi: 10.55123/storage.v3i2.3602.

[8] T. Wiratama Putra, A. Triayudi, and Andrianingsih, "Analisis Sentimen Pembelajaran Daring menggunakan Metode Naïve Bayes, KNN, dan Decision Tree," Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), vol. 6, no. 1, pp. 20–26, 2022, doi: 10.35870/jtik.v6i1.368.

This work is licensed under a Creative Commons Attribution 4.0 International License.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1111-1122
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/ 10.62712/juktisi.v5i1.1294

[9] A. Pinkan M. et al., "Perbandingan Metode Naïve Bayes, Decision Tree, dan KNN dalam Analisis Sentimen Aplikasi Gojek di Playstore," ZONAsi: Jurnal Sistem Informasi, vol. 7, no. 2, pp. 725–734, 2025, doi: 10.31849/zn.v7i2.26566.

[10] A. Syukron, Sardiarinto, E. Saputro, dan P. Widodo, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," Jurnal Teknologi Informasi dan Terapan (J-TIT), vol. 10, no. 1, pp. 47–50, 2023, doi: 10.25047/jtit.v10i1.313.

Copyright © 2023 **Author**, Page 1122
This work is licensed under a Creative Commons Attribution 4.0 International License.