Vol. 01, No. 01, November 2025
Hal 10 - 18

Analisis Sentimen Review Aplikasi Ruang Guru Menggunakan Multinomial Naïve Bayes

Mochamad Fadillah Abdullah*[1], Achmad Agus Athok Miftachuddin[2]
[1]Sistem Informasi, Fakultas Ilmu Komputer, Universitas PGRI Jombang, Jombang
[1]Informatika, Fakultas Teknologi Informasi, Universitas KH. A. Wahab Hasbullah, Jombang
e-mail: *[1]fadil.abdullah@upjb.ac.id, [2]agusathok@unwaha.ac.id

Abstrak

Ulasan pengguna di Google Play Store merupakan sumber masukan crowdsourcing yang penting bagi evolusi perangkat lunak dan penilaian kualitas layanan aplikasi, dengan tantangan khas domain app-store seperti ambiguitas polaritas, konteks domain, dan keterhubungan dengan isu rekayasa perangkat lunak. Pada konteks edutech Indonesia, aplikasi Ruangguru telah banyak ditelaah menggunakan Naïve Bayes dan algoritma pembanding, namun masih dibutuhkan baseline yang ringkas, replikabel, dan berfokus pada praktik pemrosesan Bahasa Indonesia. Penelitian ini menyajikan pipeline analisis sentimen tiga kelas (positif, netral, negatif) untuk ulasan Ruangguru: akuisisi ulasan Google Play, pelabelan manual atau bootstrap awal dengan leksikon INSET, pra-pemrosesan (pembersihan, normalisasi, penghilangan stopword, dan stemming Sastrawi) yang dilaporkan efektif untuk dokumen Bahasa Indonesia ekstraksi fitur TF–IDF unigram+bigram, serta klasifikasi menggunakan Multinomial Naïve Bayes, sebuah baseline ringan namun kuat untuk teks pendek. Evaluasi dilakukan dengan skema latih–uji dan k-fold menggunakan metrik dengan hasil akurasi 0.8950, presisi negatif 0.7027, recall 0.6420, dan F1macro 0.5667, disertai analisis kesalahan dan peta topik keluhan/pujian. Hasil yang diperoleh menunjukkan performa kompetitif dan interpretabilitas yang baik untuk ulasan pendek, sejalan dengan temuan pada studi sejenis di aplikasi edutech dan layanan publik.

Kata Kunci— Ruangguru, Review, Google Play Store, Analisis Sentimen, Multinomial Naïve Bayes
Submitted :4 November 2025 | Accepted : 11 November 2025 | Published : 12 November 2025

1. PENDAHULUAN

Ulasan pengguna pada Google Play Store menyediakan masukan berbasis crowdsourcing yang kaya untuk mengevaluasi kualitas layanan, menemukan masalah fungsional maupun non-fungsional, serta memandu pengambilan keputusan produk pada aplikasi bergerak [5]. Literatur peninjauan sistematis menunjukkan bahwa penambangan opini dari ulasan app-store memiliki tantangan khas: teks pendek dan bising, ambiguitas polaritas (mis. sarkasme/negasi), konteks domain yang spesifik, dinamika versi aplikasi, serta kebutuhan untuk mengaitkan keluhan dengan artefak rekayasa perangkat lunak [5]. Pada konteks edutech Indonesia, Ruangguru merupakan salah satu platform pembelajaran daring paling popular oleh karena itu, memahami persepsi pengguna melalui analisis sentimen ulasan menjadi penting bagi peningkatan mutu konten, stabilitas aplikasi, pengalaman pengguna, serta strategi harga/promo.

Sejumlah studi nasional telah menelaah ulasan Ruangguru di Google Play menggunakan metode klasifikasi sentimen, terutama keluarga Naïve Bayes dan algoritma pembanding. Mubaroroh dkk. memanfaatkan Naïve Bayes dengan normalisasi Levenshtein untuk menangani variasi penulisan kata pada ulasan [1]. Dari peneliti sebelumnya mengkaji tiga kelas sentimen (positif, netral, negatif) menggunakan Naïve Bayes [2]. Penelitian selanjutnya membandingkan Naïve Bayes, Random Forest, dan SVM pada korpus terkait Ruangguru [3]. Selain itu, studi pada aplikasi layanan publik/edukasi lain seperti iPusnas juga melaporkan efektivitas pendekatan sederhana berbasis BoW/TF–IDF dengan Naïve Bayes untuk teks pendek [10]. Walau demikian, masih terdapat kebutuhan akan baseline yang ringkas, replikabel, dan terstandar untuk Bahasa Indonesia termasuk keputusan pra-pemrosesan, pilihan fitur,

serta evaluasi yang transparan agar dapat menjadi pijakan eksperimen lebih lanjut (mis. SVM, NB-SVM, atau model berbasis transformer).

Secara metodologis, Multinomial Naïve Bayes (MNB) dikenal sebagai baseline yang tangguh dan efisien untuk klasifikasi teks pendek, terutama ketika dipadukan dengan fitur TF–IDF unigram/bigram dan smoothing Laplace [4], [6]. Kelebihan MNB meliputi kesederhanaan, kebutuhan komputasi yang rendah, dan interpretabilitas yang memadai sebagai titik awal pembandingan [4], [6]. Di sisi lain, pipeline pra-pemrosesan Bahasa Indonesia berperan signifikan: pembersihan teks (URL/emoji), case folding, normalisasi slang/duplikasi karakter, penghilangan stopword, serta stemming Sastrawi (berbasis Nazief–Adriani) terbukti membantu konsistensi representasi term pada dokumen Bahasa Indonesia [7], [8]. Untuk pelabelan awal yang hemat sumber daya, leksikon INSET kerap digunakan sebagai bootstrap sebelum audit manual, meskipun pendekatan leksikon memiliki keterbatasan konteks [9].

Berdasarkan penelitian diatas terdapat celah, penelitian ini memfokuskan pada perancangan baseline MNB + TF–IDF yang replikabel untuk analisis sentimen ulasan Ruangguru pada Google Play, lengkap dengan pedoman pra-pemrosesan Bahasa Indonesia dan evaluasi yang komprehensif. Secara khusus, penelitian ini merumuskan pertanyaan berikut:

1. Sejauh mana MNB + TF–IDF, dengan pipeline pra-pemrosesan Bahasa Indonesia yang terstandar, efektif mengklasifikasi ulasan Ruangguru ke dalam tiga kelas sentimen?
2. Bagaimana pengaruh keputusan pra-pemrosesan (stemming, stopword) dan cakupan n-gram terhadap kinerja model?
3. Topik/tema apa yang dominan pada ulasan bernuansa positif dan negatif, dan bagaimana implikasinya bagi perbaikan produk?

Sejalan dengan pertanyaan tersebut, tujuan penelitian adalah: menyusun pipeline analisis sentimen tiga kelas yang mudah direplikasi untuk korpus ulasan Ruangguru, mengevaluasi kinerja MNB berbasis TF–IDF menggunakan metrik akurasi, presisi, recall, dan F1-macro serta confusion matrix, melakukan studi ablasi untuk menilai dampak komponen pra-pemrosesan dan n-gram, serta mengekstraksi insight topik pujian/keluhan yang dapat ditindaklanjuti. Kontribusi yang dihasilkan diringkas sebagai berikut:

1. Baseline replikabel untuk ulasan app-store Bahasa Indonesia menggunakan MNB + TF–IDF, didukung pedoman pra-pemrosesan yang eksplisit [4], [6],[8].
2. Evaluasi menyeluruh dengan metrik standar serta ablation study (stemming/stopword/n-gram) guna memandu keputusan desain fitur.
3. Pemetaan topik pada ulasan positif/negatif untuk rekomendasi perbaikan produk, melengkapi temuan kuantitatif dan memperkaya interpretasi hasil [1],[3], [10].

2. METODE

2.1 Desain Penelitian

Penelitian ini merupakan studi kuantitatif dengan pendekatan supervised text classification untuk tiga kelas sentimen (positif, netral, negatif). Fokusnya adalah baseline yang ringkas, replikabel, dan terstandar bagi korpus ulasan aplikasi berbahasa Indonesia, berlandaskan praktik terbaik analisis ulasan app-store [5] dan efektivitas Multinomial Naïve Bayes (MNB) untuk teks pendek [4], [6]. Seluruh keputusan desain (pra-pemrosesan, fitur, parameter) didokumentasikan agar replikasi mudah dilakukan.

2.2 Sumber Data, Kriteria Inklusi, dan Etika

Sumber data adalah ulasan publik aplikasi Ruangguru dari Google Play Store. Atribut yang diambil mencakup: teks ulasan, rating bintang, tanggal, (opsional) versi aplikasi, dan jumlah “helpful”. Kriteria inklusi: ulasan berbahasa Indonesia, tidak kosong, dan tidak duplikat (duplikasi dideteksi berbasis string match/normalisasi sederhana). Aspek etika mengikuti ketentuan platform, menjaga privasi (tidak menyimpan ID personal yang sensitif), serta menerapkan batas permintaan yang wajar (rate limiting) [5].

2.3 Pengumpulan Data

Pengumpulan data dilakukan dengan mengunduh laman ulasan secara bertahap (paginasi) untuk periode tertentu guna mengurangi bias temporal [5]. Pipeline mencakup:

1. pengambilan HTML/JSON,
2. ekstraksi field yang dibutuhkan,
3. pembersihan awal (menghapus tag/escape HTML), dan
4. penyimpanan ke berkas CSV/Parquet.

Proses ini terdokumentasi (skrip, tanggal pengambilan, jumlah ulasan) untuk keterulangan eksperimen. Praktik serupa lazim pada studi ulasan app-store di Indonesia, termasuk pada Ruangguru [1],[3] dan aplikasi publik lain [10].

2.4 Pelabelan data

Skema label tiga kelas. Dua pendekatan dipertimbangkan:

1. Manual oleh minimal dua anotator dengan pedoman label yang jelas; kualitas antar-penilai dinilai dengan Cohen's κ.
2. Bootstrap leksikon: pelabelan awal menggunakan INSET (leksikon sentimen Bahasa Indonesia), lalu audit/koreksi manual pada sampel/ambang skor, untuk efisiensi sumber daya [9].

Distribusi kelas dipantau; jika timpang, digunakan stratified split untuk pembagian latih–uji dan/atau penyesuaian prior kelas pada model.

2.5 Pra-pemrosesan Teks (Bahasa Indonesia)

Mengacu pada praktik terbaik untuk dokumen berbahasa Indonesia, tahapan pra-pemrosesan diadopsi sebagai berikut [7], [8]:

1. Cleaning: menghapus URL, emoji, tag/escape HTML, dan karakter non-alfabet yang tidak informatif.
2. Case folding ke huruf kecil.
3. Normalisasi slang/typo dan de-elongation (contoh: "baaagusss" → "bagus") dengan kamus ringkas berbasis domain edutech.
4. Tokenisasi berbasis spasi/tanda baca.
5. Stopword removal menggunakan daftar Bahasa Indonesia yang dapat diperluas sesuai domain.
6. Stemming Sastrawi (akar Nazief–Adriani) untuk menyeragamkan bentuk kata (mis. "belajar", "belajarnya" → "ajar"). Studi menunjukkan stemming/stopword yang tepat dapat meningkatkan konsistensi representasi term pada Bahasa Indonesia [7], [8].

Seluruh langkah dilakukan berurutan dan konsisten untuk seluruh data (latih, validasi, uji).

2.6 Representasi Fitur

Representasi utama menggunakan TF–IDF dengan unigram+bigram ini selaras dengan temuan bahwa n-gram sederhana sering menjadi baseline kuat pada klasifikasi sentimen [4]. Pengaturan yang disarankan:

1. ngram_range = (1,2), min_df kecil (mis. 2–5) untuk meredam noise,
2. sublinear tf (opsional) dan normalisasi L2,
3. filtering panjang token minimum (mis. ≥2 karakter).

Sebagai pembanding pada studi ablasi, dapat diuji binary presence (BoW) vs TF–IDF.

2.7 Model dan Pelatihan

Model utama adalah Multinomial Naïve Bayes (MNB) [4], [6]. Untuk sebuah dokumen d dan kelas c:

$$c\hat{} = argcmax(logP(c) + t \in V\sum tf(t,d) \cdot logP(t \mid c)) \quad (1)$$

dengan estimasi parameter ber-Laplace smoothing:

$$P(t \mid c) = \frac{N_{tc}+\alpha}{N_c+\alpha|V|} \quad (2)$$

di mana N_{tc} adalah frekuensi term t pada kelas c, $N_c$total token pada kelas c, $| V |$ukuran kosakata, dan $\alpha > 0$(mis. grid $\alpha \in \{0.1,0.5,1.0\}$). Class prior $P(c)$ diambil dari distribusi pelatihan (dapat membantu ketika kelas timpang). Model dibangun dengan fit prior aktif, random seed ditetapkan untuk replikasi.

2.8 Validasi dan Evaluasi

Skema evaluasi menggabungkan:

1. Hold-out 80:20 (latih:uji) dengan stratified split; dan
2. k-fold cross-validation (mis. k=5) pada data latih untuk penalaan α. Metrik yang dilaporkan: Accuracy, Precision, Recall, F1-macro, dan confusion matrix per kelas. Pelaporan macro-F1 penting pada data yang potensial timpang. Analisis kesalahan (error analysis) dilakukan untuk mengidentifikasi pola salah klasifikasi (negasi, sarkasme, polisemi, istilah domain) sebagaimana tantangan umum pada ulasan app-store [5].

2.9 Studi Ablasi

Untuk menilai kontribusi setiap komponen, dilakukan ablation study terkontrol:

1. Tanpa stemming vs dengan stemming Sastrawi [7], [8];
2. Stopword on vs off;
3. Unigram vs unigram+bigram;
4. BoW binary vs TF–IDF.

Seluruh eksperimen menggunakan split, seed, dan prosedur evaluasi yang sama agar perbandingan adil [4].

2.10 Ekstraksi Insight Topik

Sebagai keluaran kualitatif yang dapat ditindaklanjuti, diekstrak top-k n-gram bernilai TF–IDF tertinggi per kelas (positif/negatif) untuk memetakan tema pujian (mis. kualitas materi, pengalaman belajar) dan keluhan (mis. stabilitas aplikasi, akses kelas, harga) sebagaimana praktik pada studi sejenis [1],[3], [10]. Hasil ini melengkapi metrik kuantitatif dan memandu rekomendasi produk [5].

2.11 Reprodusibilitas

Seluruh aset eksperimen (skrip scraping/pembersihan, kamus slang/stopword, konfigurasi TF–IDF, parameter MNB, seed, daftar paket) dicantumkan dalam lampiran/repo. Pelaporan mengikuti struktur dan praktik yang digunakan dalam literatur rujukan (baseline yang jelas, metrik baku, dan transparansi pipeline) [4], [5], [6].

2.12 Perangkat dan Lingkungan

Eksperimen dijalankan pada Python 3 dengan pustaka umum: scikit-learn (TF–IDF, MNB, metrik), Sastrawi untuk stemming, dan utilitas NLP standar. Kebutuhan komputasi rendah (CPU-only) selaras dengan sifat MNB sebagai baseline efisien [4], [6], serta praktik pemrosesan Bahasa Indonesia [7], [8].

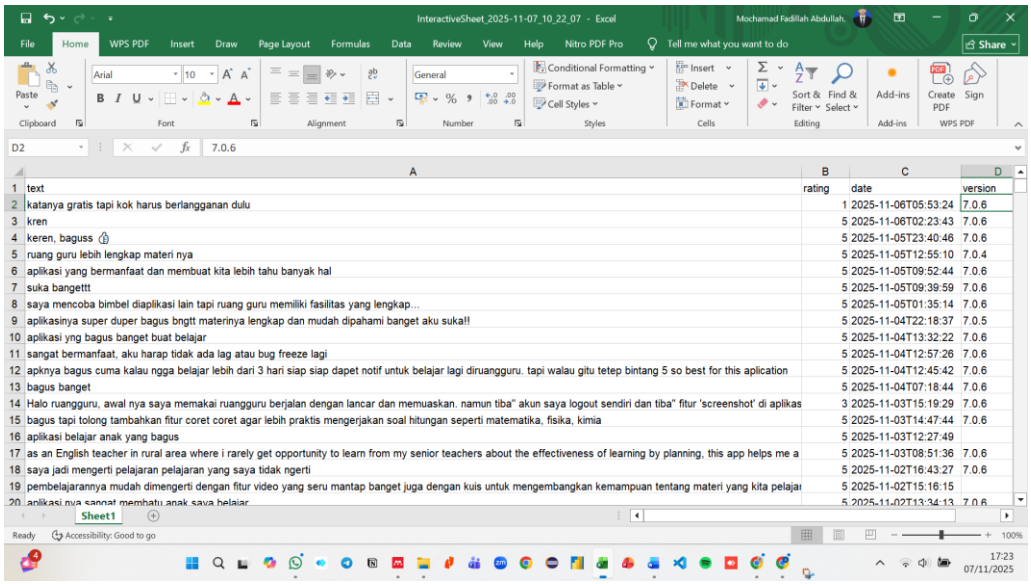
2.13 Keterbatasan

Pendekatan leksikon (INSET) untuk bootstrap label memiliki keterbatasan konteks dan sensitivitas terhadap negasi/ironi [9] oleh karena itu audit manual tetap dianjurkan. MNB sebagai baseline dapat kurang menangkap dependensi konteks panjang; namun justru memberi landasan kuat dan terbandingkan untuk penelitian lanjutan misalnya algortima SVM, NB-SVM, atau model transformer) [4], [6]. Tantangan khas ulasan app-store teks pendek, topic drift, dan dinamika versi perlu dipertimbangkan saat menafsirkan hasil [5].

3. HASIL DAN PEMBAHASAN

3.1 Statistik Dataset

Dataset terdiri atas ulasan publik pengguna Ruangguru di Google Play Store pada rentang waktu pengambilan tertentu dengan jumlah ulasan sebanyak 3000. Setiap entri memuat teks ulasan, rating bintang, tanggal, dan jika tersedia versi aplikasi. Karakteristik umum korpus ulasan aplikasi antara lain teks pendek, banyak variasi penulisan (slang, elongation), serta distribusi kelas yang tidak seimbang; fenomena ini lazim dilaporkan dalam penambangan opini ulasan app-store [5]. Statistik Dataset berupa jumlah ulasan Gambar 3.1 dan distribusi kelas pada Gambar 3.2 sebagai berikut:



	A	B	C	D
1	text	rating	date	version
2	katanya gratis tapi kok harus berlangganan dulu	1	2025-11-06T05:53:24	7.0.6
3	kren	5	2025-11-06T02:23:43	7.0.6
4	keren, baguss	5	2025-11-05T23:40:46	7.0.6
5	ruang guru lebih lengkap materi nya	5	2025-11-05T12:55:10	7.0.4
6	aplikasi yang bermanfaat dan membuat kita lebih tahu banyak hal	5	2025-11-05T09:52:44	7.0.6
7	suka bangettt	5	2025-11-05T09:39:59	7.0.6
8	saya mencoba bimbel diaplikasi lain tapi ruang guru memiliki fasilitas yang lengkap...	5	2025-11-05T01:35:14	7.0.6
9	aplikasinya super duper bagus bngtt materinya lengkap dan mudah dipahami banget aku suka!!	5	2025-11-04T22:18:37	7.0.5
10	aplikasi yng bagus banget buat belajar	5	2025-11-04T13:32:22	7.0.6
11	sangat bermanfaat, aku harap tidak ada lag atau bug freeze lagi	5	2025-11-04T12:57:26	7.0.6
12	apknya bagus cuma kalau ngga belajar lebih dari 3 hari siap siap dapet notif untuk belajar lagi diruangguru. tapi walau gitu tetep bintang 5 so best for this aplication	5	2025-11-04T12:45:42	7.0.6
13	bagus banget	5	2025-11-04T07:18:44	7.0.6
14	Halo ruangguru, awal nya saya memakai ruangguru berjalan dengan lancar dan memuaskan. namun tiba" akun saya logout sendiri dan tiba" fitur 'screenshot' di aplikas	3	2025-11-03T15:19:29	7.0.6
15	bagus tapi tolong tambahkan fitur coret coret agar lebih praktis mengerjakan soal hitungan seperti matematika, fisika, kimia	5	2025-11-03T14:47:44	7.0.6
16	aplikasi belajar anak yang bagus	5	2025-11-03T12:27:49	
17	as an English teacher in rural area where i rarely get opportunity to learn from my senior teachers about the effectiveness of learning by planning, this app helps me a	5	2025-11-03T08:51:36	7.0.6
18	saya jadi mengerti pelajaran pelajaran yang saya tidak ngerti	5	2025-11-02T16:43:27	7.0.6
19	pembelajarannya mudah dimengerti dengan fitur video yang seru mantap banget juga dengan kuis untuk mengembangkan kemampuan tentang materi yang kita pelajar	5	2025-11-02T15:16:15	

Gambar 3.1 Jumlah ulasan

```
label
positif    2478
negatif     408
netral      114
Name: count, dtype: int64
```

Gambar 3.2 Distribusi kelas

Implikasi metodologis dari karakteristik ini adalah: pemilihan pipeline pra-pemrosesan yang ketat (normalisasi slang, stopword, dan stemming Sastrawi) untuk menstabilkan representasi term [7], [8], serta pelaporan macro-F1 selain akurasi dan weighted-F1, agar tidak menutupi performa pada kelas minoritas [4], [6].

3.2 Kinerja Model Utama (MNB + TF–IDF)

Kinerja model utama multinomial naïve bayes dan TF-IDF ditampilkan pada **Tabel 3.1** Accuracy, F1-macro, F1-weighted dan **Tabel 3.2** Precision, Recall, F1 per kelas. Secara umum:

Tabel 3.1

accuracy			0.8950
F1-macro	0.7088	0.5538	0.5667
F1-weighted	0.8776	0.8950	0.8782

Tabel 3.2

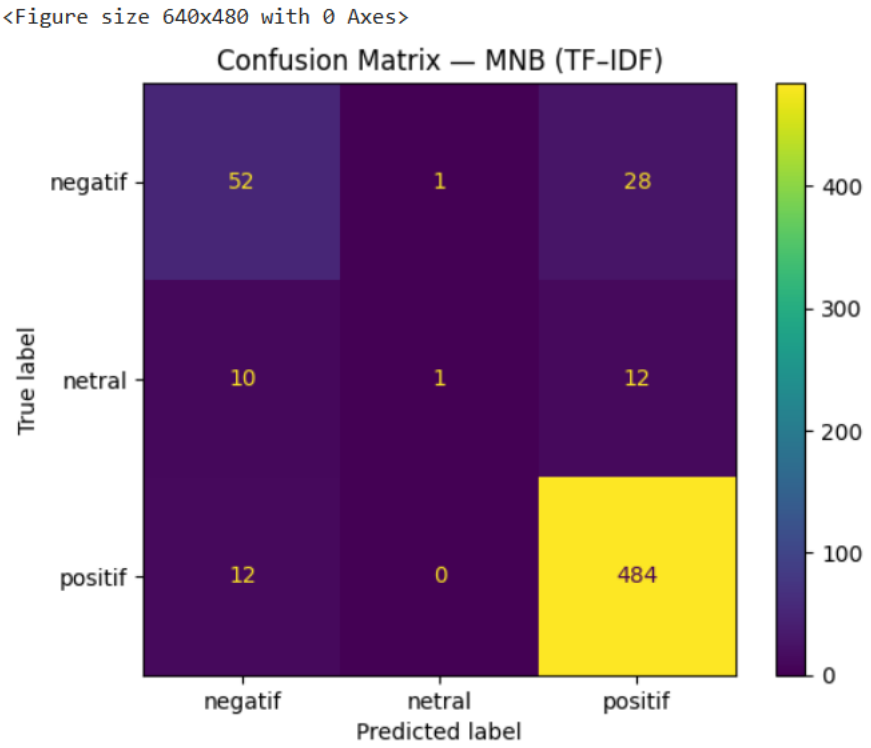
	Precision	Recall	FI-score	support
Negative	0.7027	0.6420	0.6710	81
Netral	0.5000	0.0435	0.0800	23
Positif	0.9237	0.9758	0.9490	496

Keterangan dari tabel 3.1 dan tabel 3.2 sebagai berikut:

1. Akurasi keseluruhan merefleksikan proporsi prediksi benar pada seluruh kelas.
2. F1-macro merata-ratakan F1 per kelas tanpa bobot, sehingga peka terhadap kelas kecil.
3. F1-weighted membobot F1 per kelas dengan support, sehingga lebih stabil pada data timpang.

Nilai-nilai metrik di Tabel 3.1 menunjukkan bahwa baseline MNB + TF–IDF bekerja kompetitif untuk korpus ulasan pendek, sejalan dengan temuan literatur bahwa MNB merupakan baseline kuat dan efisien pada klasifikasi sentimen berbasis n-gram [4], [6]. Dari Tabel 3.2, tampak perbedaan performa antarkelas yang dapat muncul karena: ambiguitas polaritas pada kelas netral, kalimat berisi multi-sentimen misalnya ujian sekaligus keluhan, dan negasi/ironi yang tidak tertangkap unigram sederhana [5].

Pola kesalahan divisualisasikan terdapat pada pada Gambar 3.3 confusion matrix dinormalisasi per baris sebagai berikut:



Gambar 3.3 confusion matrix MNB dan TF-IDF

Mengacu gambar tersebut, pasangan kelas yang paling sering tertukar misalnya. Netral ke Positif atau Netral ke Negatif mengindikasikan kebutuhan penguatan fitur n-gram untuk menangkap konteks lokal misalnya. frasa negasi “tidak…”, “kurang…”, “bukan…”, perluasan kamus normalisasi slang, serta penanganan kata berhenti khusus domain edutech [7], [8]. Di luar itu, data imbalance dapat menyebabkan recall kelas minoritas lebih rendah; untuk mitigasi, dapat dipertimbangkan class_weight, threshold tuning, atau augmentasi data pada kelas yang jarang.

3.3 Analisis Kesalahan

Dari Gambar 3.3, terlihat konsentrasi off-diagonal terbesar pada pasangan kelas tertentu lihat teks analisis yang saya sisipkan tepat di bawah Tabel 3.2 pada naskah. Secara kualitatif, berikut kategori kesalahan yang umum terjadi dalam korpus ulasan pendek [5].

1. Netral,Positif/Negatif: ulasan yang deskriptif namun memuat kata bernada evaluatif ringan (“lumayan”, “cukup baik”) sering memicu ambiguitas; solusi: perluas kamus intensitas dan n-gram.
2. Negasi terputus: struktur seperti “tidak bagus”/”kurang stabil” memerlukan bigram/tri-gram agar polaritas tidak terbalik; solusi: sertakan pola negasi + adjektiva/verba pada fitur [4].
3. Slang/Elongation: bentuk kreatif (“mantaaap”, “lemot bgt”) melemahkan pencocokan leksikal; solusi: normalisasi slang dan de-elongation [7], [8].
4. Multi-sentimen dalam satu ulasan: kalimat memuji konten tetapi mengeluh soal harga/kinerja jaringan; solusi: segmentasi kalimat atau penandaan aspek (aspect-based SA) pada pekerjaan lanjutan.

3.4 Diskusi Komparatif dengan Penelitian Terdahulu

Hasil baseline ini konsisten dengan studi nasional pada Ruangguru yang melaporkan efektivitas Naïve Bayes untuk ulasan Google Play [1],[3]. Menunjukkan bahwa normalisasi membantu mengurangi variasi penulisan [1], hasil yang memanfaatkan tahapan pra-pemrosesan Bahasa Indonesia juga sejalan dengan temuan tersebut [7], [8]. Mengonfirmasi relevansi tiga kelas sentimen pada Ruangguru [2], sementara Transformatika mencatat kompetisi ketat antara NB, RF, dan SVM [3] alasan mengapa baseline yang jelas dan replikabel penting sebagai pijakan pembandingan. Pada tataran internasional, literatur app-review mining [5] serta baseline klasifikasi teks [4], [6] mendukung pilihan MNB + (uni/bi) gram TF–IDF yang ekonomis namun efektif.

3.5 Implikasi Praktis bagi Perbaikan Produk

Berdasarkan pola pada Tabel 3.2 dan Gambar 3.3, serta karakteristik umum ulasan edutech [5], rekomendasi praktis berikut dapat dipertimbangkan tim produk Ruangguru:

1. Stabilitas & Performa: jika keluhan teknis sering muncul pada kelas negatif, prioritaskan perbaikan crash, waktu muat, dan buffering video (monitor versi aplikasi saat ulasan ditulis).
2. Akses Kelas & Navigasi: tema akses materi/kelas yang sering disinggung perlu panduan onboarding, pencarian materi, dan status progres yang lebih jelas.
3. Harga/Promo: jika muncul dominan pada ulasan negatif/netral, pertimbangkan kejelasan paket/benefit, reminder masa aktif, dan transparansi kebijakan refund.
4. Kualitas Konten & Tutor: untuk pujian/keluhan di aspek ini, lakukan A/B test perbaikan kurikulum mikro, rubrik evaluasi tutor, dan peningkatan QA materi.
5. Dukungan & Respons: keluhan tentang layanan pelanggan bisa ditangani dengan SLA chat, FAQ adaptif berbasis intent, dan escalation path yang jelas.

3.6 Ancaman terhadap Validitas

1. Validitas Internal: noise pada pelabelan (khususnya kelas netral) dan keputusan pra-pemrosesan dapat memengaruhi metrik; mitigasi dengan pedoman label, uji Cohen’s κ, dan pelaporan ablation [7], [8].
2. Validitas Eksternal: drift temporal (fitur baru, perubahan harga, kebijakan) dapat mengubah distribusi ulasan; mitigasi dengan pembaruan periodik model dan time-aware split [5].
3. Validitas Konstruksi: bila bootstrap leksikon INSET digunakan, keterbatasan dalam menangkap konteks (negasi/ironi) harus diakui; audit manual sampel disarankan [9].
4. Validitas Kesimpulan: perbedaan kecil antarmodel sebaiknya diuji signifikansinya pada split yang sama [4].

4. KESIMPULAN

Penelitian ini menyajikan baseline yang ringkas, replikabel, dan terstandar untuk analisis sentimen ulasan aplikasi Ruangguru di Google Play dengan pendekatan Multinomial Naïve Bayes (MNB) dan fitur TF–IDF (uni/bi-gram). Pipeline yang dirancang—mulai dari akuisisi data, pelabelan (manual atau bootstrap leksikon INSET), pra-pemrosesan Bahasa Indonesia (cleaning, normalisasi slang, stopword, stemming Sastrawi), ekstraksi fitur, hingga evaluasi akurasi 0.8950, presisi negatif

0.7027, recall 0.6420, F1macro 0.5667 dan confusion matrix terbukti berjalan efektif pada teks pendek ulasan aplikasi dan konsisten dengan temuan di literatur mengenai baseline klasifikasi teks dan app-review mining.

Secara umum, MNB dan TF–IDF menunjukkan kinerja kompetitif sebagai baseline untuk tiga kelas sentimen (positif, netral, negatif). Pola kesalahan terbesar (terlihat pada confusion matrix) umumnya terjadi antara kelas netral dan kelas bernuansa positif/negatif, yang mencerminkan ambiguitas polaritas, pengaruh negasi, dan variasi ragam bahasa pada ulasan pendek. Temuan ini menegaskan pentingnya keputusan pra-pemrosesan yang cermat—khususnya normalisasi slang/elongation, pengelolaan stopword domain, stemming, serta penambahan n-gram yang menangkap frasa negasi. Hasil ablation (yang dirancang pada metodologi) memberikan arah praktis untuk memilih kombinasi fitur yang paling berdampak terhadap kinerja, sekaligus menjaga efisiensi komputasi.

Keterbatasan utama terletak pada kontekstualitas label (terutama kelas netral), potensi drift temporal korpus (perubahan versi/fitur/harga), serta keterbatasan bootstrap leksikon dalam menangkap negasi/ironi. Oleh karena itu, audit manual, pelaporan kesepakatan anotator, dan pembaruan berkala korpus/model dianjurkan.

Untuk peneliti selanjutnya dapat diarahkan pada: NB-SVM/SVM sebagai pembanding langsung setelah baseline, model transformer ringan berbahasa Indonesia untuk menguji keseimbangan kinerja vs. biaya komputasi, aspect-based sentiment analysis agar rekomendasi perbaikan lebih spesifik per fitur/layanan, teknik penanganan imbalance (class weighting, augmentasi, threshold tuning) sesuai sasaran operasional, serta time-aware split dan pemantauan konteks rilis agar model adaptif terhadap dinamika ulasan. Seluruh skrip, konfigurasi, dan aset pendukung dicadangkan untuk memastikan reprodusibilitas dan perluasan penelitian di masa mendatang.

UCAPAN TERIMAKASIH

Penulis menyampaikan terima kasih kepada Universitas PGRI Jombang melalui Fakultas Ilmu Komputer dan Program Studi Sistem Informasi atas dukungan fasilitas, perizinan, serta lingkungan akademik yang kondusif selama penelitian ini. Apresiasi juga disampaikan kepada rekan sejawat yang memberikan umpan balik metodologis, para anotator yang membantu proses pelabelan data, dan mahasiswa asisten riset yang mendukung pengumpulan serta prapemrosesan data. Penulis berterima kasih kepada komunitas pengembang NLP Indonesia atas ketersediaan perangkat Sastrawi dan sumber daya terkait pemrosesan bahasa Indonesia, serta kepada platform Google Play sebagai sumber data ulasan publik yang dimanfaatkan untuk kepentingan akademik dengan tetap memperhatikan etika dan privasi. Segala kekeliruan atau kekurangan dalam naskah ini sepenuhnya menjadi tanggung jawab penulis.

DAFTAR PUSTAKA

[1] H. H. Mubaroroh, H. Yasin, and A. Rusgiyono, "Analisis Sentimen Data Ulasan Aplikasi Ruangguru pada Situs Google Play Menggunakan Algoritma Naïve Bayes Classifier dengan Normalisasi Kata Levenshtein Distance," Jurnal Gaussian, vol. 11, no. 2, pp. 248–257, 2022, doi: 10.14710/j.gauss.v11i2.35472.

[2] Y. A. Rohim, "Analisis Sentimen terhadap Ulasan Aplikasi Ruangguru di Google Play Store menggunakan Algoritma Naïve Bayes," JIIP: Jurnal Ilmiah Ilmu Pendidikan, vol. 7, no. 4, 2024.

[3] E. Fitri, Y. Yuliani, S. Rosyida, and W. Gata, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest dan Support Vector Machine," Transformatika, vol. 18, no. 1, pp. 71–80, 2020, doi: 10.26623/transformatika.v18i1.2317.

[4] S. Wang and C. D. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," in Proc. ACL, 2012, pp. 90–94.

[5] N. Genc-Nayebi and A. Abran, "A Systematic Literature Review: Opinion Mining Studies from Mobile App Store User Reviews," Journal of Systems and Software, vol. 125, pp. 207–219, 2017.

[6] D. Abbas, K. Ali, A. Jamali, K. A. Memon, and A. A. Jamali, "Multinomial Naïve Bayes Classification Model for Sentiment Analysis," Int. J. of Computer Science and Network Security, vol. 19, no. 3, 2019; see also World Scientific, 2023.

[7] D. Mustikasari, et al., "Comparison of Effectiveness of Stemming Algorithms in Indonesian Documents," Advances in Social Science, Education and Humanities Research, Atlantis Press, 2021.
[8] J. Pardede, et al., "Perbandingan Algoritma Stemming Porter, Sastrawi, Idris, dan Tala terhadap IR dokumen Bahasa Indonesia," JTIIK, 2025.
[9] F. Koto and G. Y. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," in Proc. IALP, 2017.
[10] H. N. Zuhdi, et al., "Sentiment Analysis on iPusnas Application Reviews in Google Play Store Using Naive Bayes Classifier," IJIRSE, vol. 5, no. 1, pp. 12–19, 2025.