

Analisis Sentimen Multikelas Bahasa Jawa: Studi Komparatif dan Ensemble Learning Model IndoBERT dan JavaBERT

Dian Nurdiana^{*1}, Septian Nuno Zildjian², Muhammad Fawaz Saputra³

¹²Program Studi Sains Data, Universitas Terbuka, Jakarta, ³Program Studi Sistem Informasi, Fakultas Teknik, Universitas Mulawarman

e-mail: ^{*1}dian.nurdiana@ecampus.ut.ac.id, ²septiannunozildjian@gmail.com, ^{3m}m_fawazsaputra@ft.unmul.ac.id

Abstrak

Bahasa Jawa merupakan bahasa daerah dengan jumlah penutur terbesar di Indonesia, namun sumber daya pemrosesan bahasa alami (Natural Language Processing, NLP) untuknya masih sangat terbatas dibandingkan Bahasa Indonesia. Penelitian ini menyajikan studi komparatif untuk tugas analisis sentimen multikelas (Positif, Netral, Negatif) pada teks berbahasa Jawa menggunakan dua model pre-trained berbasis Transformer, yaitu IndoBERT (indobert-base-p1) dan JavaBERT (javanese-bert-small), serta model ensemble soft-voting dari keduanya. Karena korpus mentah (200.000 kalimat) tidak memiliki label, penelitian ini menerapkan pelabelan otomatis berbasis leksikon sentimen Bahasa Jawa yang disusun manual dengan penanganan negasi dan kalimat kontras. Ketiga model di-fine-tune dan dievaluasi pada data uji yang sama (20.000 kalimat), menghasilkan akurasi, presisi, recall, dan F1-score sebesar 1,00 secara konsisten. Hasil yang sangat tinggi ini dibahas secara kritis sebagai konsekuensi langsung dari kuatnya korelasi leksikal antara skema pelabelan otomatis dan fitur permukaan teks, bukan sebagai bukti pemahaman semantik mendalam. Untuk menguji generalisasi di luar distribusi latih, dilakukan pengujian kualitatif tambahan pada kalimat berdialek Ngapak (Banyumasan) yang tidak terwakili dalam leksikon maupun data latih. Pada kalimat berdialek standar, ketiga model konsisten sepakat; namun pada kalimat berdialek Ngapak, IndoBERT dan JavaBERT saling berbeda pendapat dan model ensemble nyaris seri (50,0% berbanding 50,0%), mengindikasikan keterbatasan generalisasi lintas-dialek. Temuan ini menegaskan perlunya benchmark berlabel manusia, cakupan dialek yang lebih luas, dan evaluasi di luar distribusi sebagai arah penelitian selanjutnya.

Kata kunci—Analisis Sentimen, Bahasa Jawa, IndoBERT, JavaBERT, Ensemble Learning

1. PENDAHULUAN

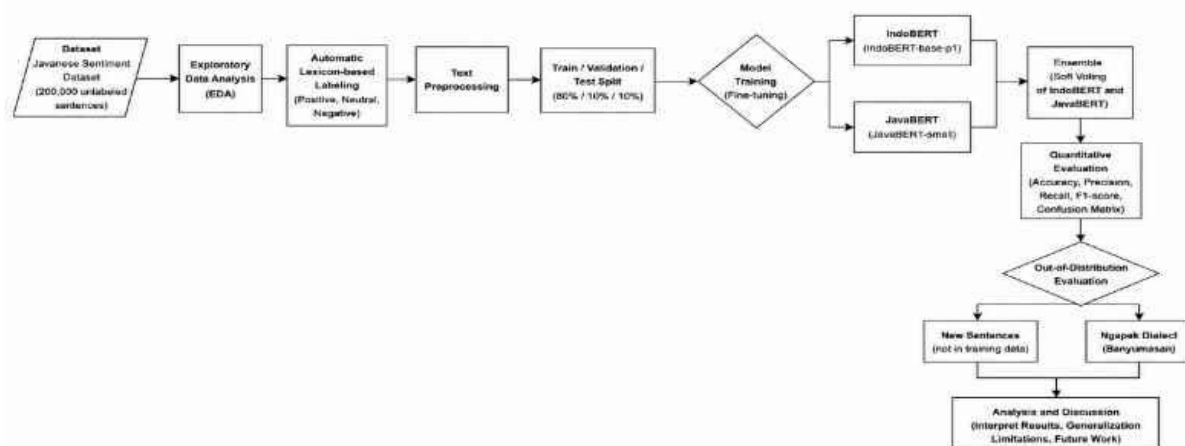
Bahasa Jawa dituturkan oleh lebih dari 80 juta orang dan merupakan bahasa daerah dengan jumlah penutur terbanyak di Asia Tenggara, namun ketersediaan sumber daya NLP untuknya masih jauh tertinggal dibandingkan Bahasa Indonesia maupun bahasa-bahasa besar dunia [4]. Ketimpangan ini menjadikan Bahasa Jawa sebagai low-resource language dalam konteks pemrosesan bahasa alami: korpus berlabel yang tersedia untuk tugas seperti analisis sentimen sangat terbatas, sementara variasi register (ngoko, krama) dan dialek regional (misalnya Ngapak/Banyumasan, Osing) menambah kompleksitas linguistik yang jarang tercakup dalam korpus yang ada. Di sisi lain, model bahasa pre-trained berbasis arsitektur Transformer [18] seperti BERT [1] telah terbukti efektif untuk berbagai tugas klasifikasi teks setelah di-fine-tune pada tugas hilir (downstream task). BERT memperkenalkan paradigma pre-training dua arah menggunakan masked language modeling yang secara signifikan meningkatkan performa berbagai tugas NLU [1]. Untuk Bahasa Indonesia, IndoBERT sebagaimana dirilis dalam kerangka kerja IndoNLU dilatih pada korpus Indo4B (sekitar 4 miliar kata) yang mencakup teks media sosial, berita, dan sumber daring lain, serta menunjukkan performa state-of-the-art pada dua belas tugas benchmark Bahasa Indonesia, sehingga menjadi tulang punggung berbagai aplikasi NLP berbahasa Indonesia [2], [16]. Untuk Bahasa Jawa, Wongso dkk. [3] merilis JavaBERT, model BERT berskala kecil yang dilatih menggunakan arsitektur causal dan masked language modeling pada korpus Wikipedia Berbahasa Jawa yang berukuran jauh lebih kecil. Winata dkk. [4] melalui NusaX menegaskan kelangkaan data berlabel paralel untuk sepuluh bahasa daerah Indonesia termasuk Jawa, dan menunjukkan bahwa performa model pada bahasa daerah umumnya masih tertinggal dibandingkan Bahasa Indonesia, sebuah kesenjangan yang juga ditegaskan pada survei tantangan NLP untuk bahasa-bahasa daerah Indonesia yang kurang terwakili [17].

Karena IndoBERT dilatih pada korpus berskala jauh lebih besar namun tidak spesifik pada Bahasa Jawa, sedangkan JavaBERT lebih kecil namun spesifik-bahasa, pertanyaan yang relevan secara praktis adalah model mana yang lebih andal untuk analisis sentimen teks Jawa, dan apakah kombinasi (ensemble) keduanya dapat memberikan hasil yang lebih baik dan lebih stabil dibandingkan model

tunggal. Penelitian ini bertujuan menjawab pertanyaan tersebut melalui studi komparatif yang sistematis. Kontribusi utama penelitian ini mencakup empat hal. Pertama, merancang dan menerapkan skema pelabelan otomatis berbasis leksikon sentimen Bahasa Jawa, dengan penanganan negasi dan kalimat kontras, untuk memberi label pada korpus mentah berskala 200.000 kalimat yang sebelumnya tidak berlabel. Kedua, melakukan fine-tuning dan evaluasi komparatif terhadap IndoBERT dan JavaBERT pada tugas analisis sentimen multikelas Bahasa Jawa, serta membangun model ensemble soft-voting berbobot dari keduanya. Ketiga, melakukan pengujian kualitatif di luar distribusi data latih menggunakan kalimat-kalimat baru, termasuk kalimat berdialek Ngapak, untuk mengevaluasi generalisasi model secara lebih realistis di luar metrik evaluasi standar. Keempat, membahas secara kritis fenomena skor evaluasi sempurna (akurasi/F1 = 1,00) yang muncul akibat pelabelan berbasis leksikon, beserta implikasinya terhadap validitas klaim performa model, sebagai bagian dari analisis keterbatasan penelitian.

2. METODE PENELITIAN

Alur kerja penelitian mencakup tujuh tahap: (1) eksplorasi data, (2) pelabelan otomatis berbasis leksikon, (3) *preprocessing* teks, (4) pembagian data, (5) *fine-tuning* IndoBERT dan JavaBERT, (6) pembentukan model *ensemble*, dan (7) pengujian kualitatif pada kalimat baru.



Gambar 1. Diagram alir metode penelitian

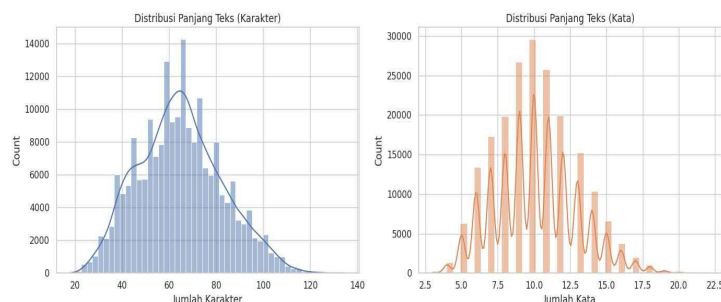
2.1 Dataset

Dataset yang digunakan adalah *Javanese Sentiment Dataset* yang bersumber dari Kaggle [7]. Data ini dihimpun dari berbagai sumber seperti komentar media sosial, ulasan Google Maps, ulasan Play Store, serta lain-lain, yang dikumpulkan dan diunggah di platform Kaggle. Dataset terdiri atas 200.000 baris teks berbahasa Jawa tanpa label sentimen. Tabel 1 merangkum statistik deskriptif panjang teks. Rata-rata panjang kalimat adalah 64,7 karakter (SD = 17,9) atau 10,1 kata (SD = 2,8), dengan rentang 3-22 kata per kalimat mengindikasikan korpus yang didominasi oleh kalimat pendek bergaya ulasan/komentar.

Tabel 1 - Statistik Deskriptif Panjang Teks pada Dataset (n = 200.000)

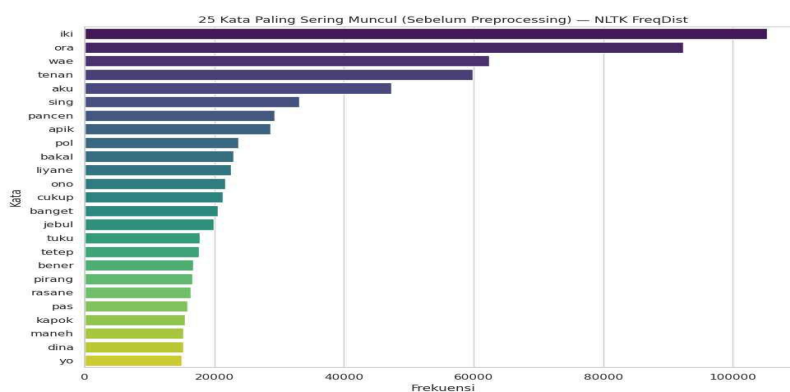
Statistik	Jumlah Karakter	Jumlah Kata
Rata-rata	64,73	10,12
Std. Deviasi	17,93	2,83
Minimum	18	3
Kuartil 1 (25%)	52	8
Median (50%)	64	10
Kuartil 3 (75%)	76	12
Maksimum	135	22

Untuk memberikan gambaran visual mengenai bentuk distribusi tersebut, Gambar 2 menampilkan histogram jumlah karakter dan jumlah kata per kalimat pada seluruh 200.000 data.



Gambar 2. Distribusi panjang teks dalam karakter (kiri) dan kata (kanan).

Gambar 2 menunjukkan distribusi panjang teks berbentuk mendekati normal dan sedikit menjulur ke kanan, konsisten dengan gaya penulisan ulasan singkat. Gambar 3 dan 4 menyajikan frekuensi kata dan *word cloud* sebelum *preprocessing*; kata fungsi seperti *iki* (ini), *ora* (tidak), *wae* (saja), dan *tenan* (sungguh) paling sering muncul, mencerminkan pola penekanan emosi yang khas dalam Bahasa Jawa informal.



Gambar 3. 25 kata paling sering muncul sebelum *preprocessing* (NLTK FreqDist).

Selain disajikan dalam bentuk diagram batang pada Gambar 3, frekuensi kemunculan kata-kata tersebut juga divisualisasikan dalam bentuk word cloud pada Gambar 4, di mana ukuran setiap kata merepresentasikan proporsi frekuensi kemunculannya secara keseluruhan pada korpus.



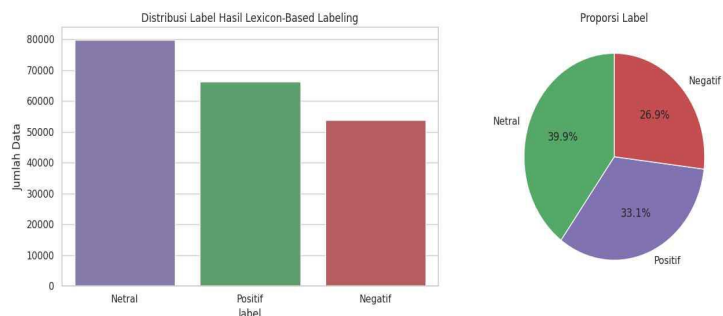
Gambar 4. *Word cloud* keseluruhan dataset sebelum *preprocessing*.

2.2 Pelabelan Otomatis Berbasis Leksikon

Karena dataset mentah tidak memiliki label, penelitian ini menyusun leksikon sentimen Bahasa Jawa secara manual yang terdiri atas daftar kata positif dan negatif beserta bobot skalarnya (misalnya *apik* = +2, *elek* = -2, *lumayan* = bobot rendah), dilengkapi mekanisme penanganan negasi (mis. *ora apik*) dan penanda kontras (*nanging*, *senadyan*) yang dapat membalik atau meredam polaritas klausa sebelumnya. Skor sentimen suatu kalimat dihitung sebagai jumlah berbobot dari kata-kata bermuatan sentimen yang terdeteksi, kemudian dipetakan ke tiga kelas menggunakan ambang batas: skor > 0 diberi label **Positif**, skor = 0 diberi label **Netral**, dan skor < 0 diberi label **Negatif**.

Hasil pelabelan menghasilkan distribusi kelas yang tidak seimbang namun tidak ekstrem: Netral 79.900 data (39,95%), Positif 66.280 data (33,14%), dan Negatif 53.820 data (26,91%), sebagaimana ditunjukkan pada Gambar 5. Gambar 6 menampilkan *word cloud* per kelas label; kata seperti *apik*, *puas*, dan

jebul (positif), *liyane* dan *ngono* (netral), serta *nyesel*, *kapok*, dan *elek* (negatif) tampak dominan pada masing-masing kelas sebuah indikasi awal bahwa label yang dihasilkan sangat konsisten secara leksikal dengan kata-kata dalam leksikon itu sendiri.



Gambar 5. Distribusi label hasil pelabelan berbasis leksikon (kiri: jumlah, kanan: proporsi).

Untuk memperjelas karakteristik kosakata pada masing-masing kelas sentimen, dilakukan pula visualisasi word cloud secara terpisah untuk setiap label sebagaimana disajikan pada Gambar 6, guna mengidentifikasi kata-kata yang paling dominan pada kelas Positif, Netral, dan Negatif.



Gambar 6. Word cloud per kelas label: Positif, Netral, dan Negatif.

2.3 Preprocessing Teks

Tahap *preprocessing* mencakup *lowercasing*, penghapusan URL/*mention*/karakter non-alfabet, serta penghapusan *stopword*. Karena NLTK tidak memiliki korpus *stopword* Bahasa Jawa, daftar *stopword* *ngoko/krama* umum disusun secara manual dan digabungkan dengan *stopword* Bahasa Indonesia dari NLTK, menghasilkan 826 *stopword* gabungan. Teks hasil pembersihan (*text_clean*) digunakan sebagai masukan model, sedangkan versi tanpa *stopword* (*text_clean_nostop*) digunakan untuk analisis eksploratif kata kunci per label.

2.4 Analisis Sentimen dan Pelabelan Berbasis Leksikon

Analisis sentimen berbasis leksikon merupakan pendekatan yang umum digunakan ketika data berlabel manusia tidak tersedia, dengan memanfaatkan daftar kata bermuatan sentimen beserta bobotnya untuk menghitung skor sentimen suatu kalimat [5], [8], [9]. Pendekatan ini praktis untuk menghasilkan label awal berskala besar secara otomatis, tetapi secara inheren membatasi keragaman sinyal yang dipelajari model hilir pada sinyal leksikal yang sama dengan yang digunakan untuk membangkitkan label sebuah isu metodologis yang menjadi salah satu fokus pembahasan kritis pada penelitian ini (lihat Bagian 3.4).

2.5 Ensemble Learning

Kombinasi beberapa model (*ensemble*) secara umum dapat meningkatkan akurasi dan stabilitas prediksi dibandingkan model tunggal, khususnya ketika model-model anggota memiliki pola kesalahan yang saling melengkapi [6]. Skema *soft voting* yang merata-ratakan probabilitas keluaran (*softmax*) dari beberapa model merupakan salah satu bentuk *ensemble* yang paling sederhana dan banyak digunakan dalam klasifikasi teks. Penelitian ini mengadopsi skema tersebut dengan pembobotan berdasarkan performa validasi masing-masing model anggota.

2.6 Pembagian Data

Data dibagi menjadi subset latih, validasi, dan uji dengan rasio 80:10:10 secara *stratified* terhadap label, menghasilkan 160.000 data latih, 20.000 data validasi, dan 20.000 data uji (Tabel 2). Proporsi kelas pada subset uji (Negatif 5.382; Netral 7.990; Positif 6.628) konsisten dengan proporsi pada keseluruhan dataset, mengonfirmasi keberhasilan proses *stratified sampling*.

Tabel 2 - Pembagian Dataset

Subset	Jumlah Data	Proporsi
Latih (Train)	160.000	80%
Validasi	20.000	10%
Uji (Test)	20.000	10%
Total	200.000	100%

2.7 Arsitektur dan Fine-Tuning Model

Dua model *pre-trained* di-*fine-tune* untuk klasifikasi sekuens tiga kelas menggunakan *head* klasifikasi standar di atas representasi token [CLS]:

1. **IndoBERT** - indobenchmark/indobert-base-p1 [2], model BERT-*base* 12-layer yang dilatih pada korpus Indo4B berbahasa Indonesia.
2. **JavaBERT** - w11wo/javanese-bert-small [3], model BERT berskala kecil yang dilatih pada korpus Wikipedia Berbahasa Jawa.

Kedua model di-*fine-tune* dengan hyperparameter yang identik agar perbandingan adil (Tabel 3), menggunakan pustaka Hugging Face Transformers Trainer dengan strategi *early stopping* berdasarkan F1 validasi, mengikuti praktik umum fine-tuning BERT untuk klasifikasi teks [19].

Tabel 3 - Hyperparameter Fine-Tuning

Parameter	Nilai
Panjang token maksimum	96
Ukuran batch	16
Jumlah epoch	3
Learning rate	2×10^{-5}
Random seed	42

Proses *fine-tuning* IndoBERT menghasilkan *training loss* akhir sebesar 0,0215 dalam waktu $\pm 96,4$ menit (15.000 *step*, ~ 83 sampel/detik), sedangkan JavaBERT menghasilkan *training loss* akhir sebesar 0,0144 dalam waktu $\pm 93,1$ menit (~ 86 sampel/detik) menunjukkan konvergensi yang sangat baik pada kedua model.

2.8 Ensemble Soft-Voting

Model *ensemble* dibentuk dengan merata-ratakan probabilitas *softmax* keluaran IndoBERT (p_I) dan JavaBERT (p_J) menggunakan bobot w_I dan w_J yang ditentukan berdasarkan F1 validasi masing-masing model:

$$p_{ensemble} = w_I \cdot p_I + w_J \cdot p_J, \quad w_I + w_J = 1 \quad (1)$$

Label akhir ditentukan melalui argmax dari $p_{ensemble}$. Karena F1 validasi kedua model identik (1,0000), bobot yang dihasilkan adalah $w_I = w_J = 0,50$ (*unweighted average*). Skema ini dipilih karena kesederhanaannya dan karena tidak memerlukan pelatihan tambahan (*meta-learner*) di luar kedua model dasar.

3. HASIL DAN PEMBAHASAN

3.1 Evaluasi Kuantitatif pada Data Uji

Tabel 4 menyajikan ringkasan *classification report* untuk ketiga model pada 20.000 data uji, dan Tabel 5 merangkum perbandingan akhir. Ketiga model IndoBERT, JavaBERT, dan *ensemble* mencapai akurasi, presisi, *recall*, dan F1-score sebesar 1,00 secara merata pada seluruh kelas, sebagaimana juga terlihat pada matriks konfusi (Gambar 7-9) yang seluruhnya diagonal sempurna tanpa kesalahan klasifikasi. Gambar 10 memvisualisasikan kesetaraan performa ketiga model pada seluruh metrik.

Tabel 4 - Classification Report pada Data Uji (20.000 sampel)

Model	Kelas	Presisi	Recall	F1
IndoBERT	Negatif	1,00	1,00	1,00
	Netral	1,00	1,00	1,00
	Positif	1,00	1,00	1,00
JavaBERT	Negatif	1,00	1,00	1,00

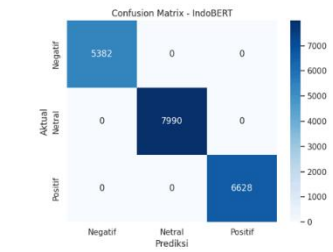
Ensemble	Netral	1,00	1,00	1,00
	Positif	1,00	1,00	1,00
	Negatif	1,00	1,00	1,00
	Netral	1,00	1,00	1,00
	Positif	1,00	1,00	1,00

Pola presisi, recall, dan F1 yang identik sebesar 1,00 pada seluruh kelas dan ketiga model (Tabel 4) tidak menyisakan variasi kesalahan klasifikasi sama sekali, sebuah kondisi yang jarang ditemukan pada tugas analisis sentimen berbasis bahasa alami dan patut dicurigai sebagai indikasi adanya sinyal permukaan yang sangat kuat pada data berlabel, sejalan dengan fenomena shortcut learning pada model berkapasitas tinggi [10] dan artefak anotasi yang telah dilaporkan pada berbagai tugas NLP berbasis pelabelan otomatis maupun crowd-sourced [11]. Interpretasi lebih rinci atas temuan ini disajikan pada Bagian 3.4.

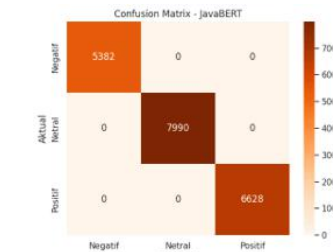
Tabel 5 - Perbandingan Performa Akhir Ketiga Model

Model	Acc.	Prec.	Recall	F1
IndoBERT	1,00	1,00	1,00	1,00
JavaBERT	1,00	1,00	1,00	1,00
Ensemble	1,00	1,00	1,00	1,00

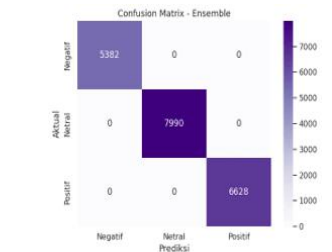
Untuk memperjelas performa klasifikasi pada tingkat kelas secara individual, matriks konfusi ketiga model divisualisasikan pada Gambar 7-9 berikut.



Gambar 7. Matriks konfusi IndoBERT pada data uji.

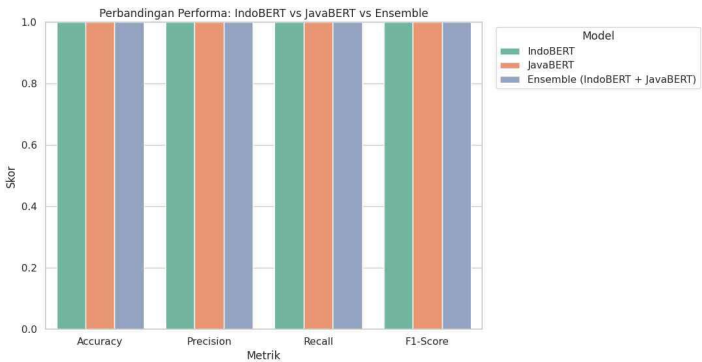


Gambar 8. Matriks konfusi JavaBERT pada data uji.



Gambar 9. Matriks konfusi model ensemble pada data uji.

Guna mempermudah perbandingan performa antar model secara ringkas, keempat metrik evaluasi pada Tabel 5 tersebut divisualisasikan bersama dalam bentuk diagram batang pada Gambar 10.



Gambar 10. Perbandingan Accuracy, Precision, Recall, dan F1-Score antara IndoBERT, JavaBERT, dan Ensemble.

Sekilas, hasil ini tampak sebagai bukti bahwa ketiga model memiliki performa sempurna. Namun sebagaimana dibahas secara rinci pada Bagian 3.4, interpretasi tersebut perlu dilakukan dengan sangat hati-hati mengingat metodologi pelabelan yang digunakan.

3.2 Pengujian Kualitatif pada Kalimat Baru (Notebook)

Sebagai pelengkap evaluasi kuantitatif, lima kalimat baru yang tidak berasal dari dataset diuji menggunakan model ensemble (Tabel 6). Model ensemble berhasil mengklasifikasikan seluruh kalimat dengan confidence 1,00, termasuk kalimat kontras seperti “*Senadyan rega rada mahal, nanging kualitase*

pancen top markotop” (Meski harganya agak mahal, tapi kualitasnya benar-benar top) yang berhasil diklasifikasikan Positif - menunjukkan bahwa mekanisme penanganan kontras pada leksikon berhasil diteruskan dengan baik ke model.

Tabel 6 - Pengujian Kualitatif pada Kalimat Baru (Model Ensemble)

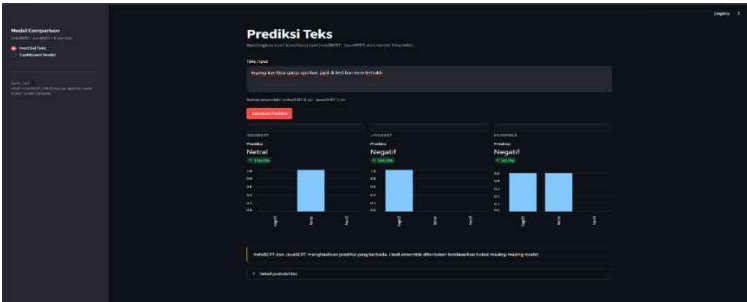
Kalimat	Prediksi	Conf.
Pelayanane apik banget, aku puas tenan tuku ing kene.	Positif	1,00
Barange elek, ora sesuai karo gambar, aku kapok.	Negatif	1,00
Regane biasa wae, ora ono sing istimewa.	Negatif	1,00
Senadyan rega rada mahal, nanging kualitase pancen top markotop.	Positif	1,00
Pengirimane lemot banget, ngecewakke pelanggan.	Negatif	1,00

Menariknya, kalimat “*Regane biasa wae, ora ono sing istimewa*” (Harganya biasa saja, tidak ada yang istimewa) yang secara intuitif lebih dekat pada sentimen netral diklasifikasikan sebagai Negatif dengan *confidence* penuh. Hal ini mengindikasikan bahwa frasa *ora ono sing istimewa* (tidak ada yang istimewa) dipelajari model sebagai penanda negasi yang kuat, konsisten dengan pola pada leksikon pelabelan, meskipun secara pragmatis maknanya lebih netral daripada negatif.

Pengujian tambahan pada kalimat yang mengandung kata makian (“*Asu, koe janjane bisa apa ora sih.*”) menghasilkan prediksi Negatif namun dengan *confidence* hanya 0,500 (probabilitas Negatif 0,5000 dan Netral 0,4999, Positif $\approx$ 0) sebuah kasus hampir seri (*near-tie*) yang mengindikasikan ketidakpastian tinggi model terhadap kalimat retorik/informal yang tidak eksplisit tercakup pola leksikonnya.

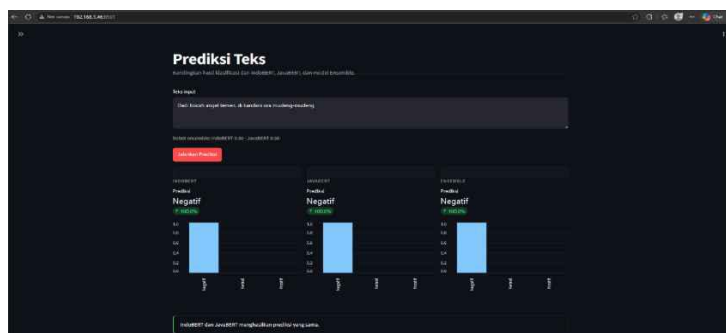
3.3 Pengujian Kualitatif pada Antarmuka Interaktif: Studi Kasus Dialek Ngapak

Untuk menguji generalisasi model secara lebih menantang, dilakukan pengujian tambahan melalui antarmuka demonstrasi interaktif (Streamlit) yang membandingkan keluaran IndoBERT, JavaBERT, dan *ensemble* secara *side-by-side*. Tiga kalimat diujikan, mencakup satu kalimat berdialek Ngapak (Banyumasan) dan dua kalimat berdialek Jawa standar (Gambar 11).



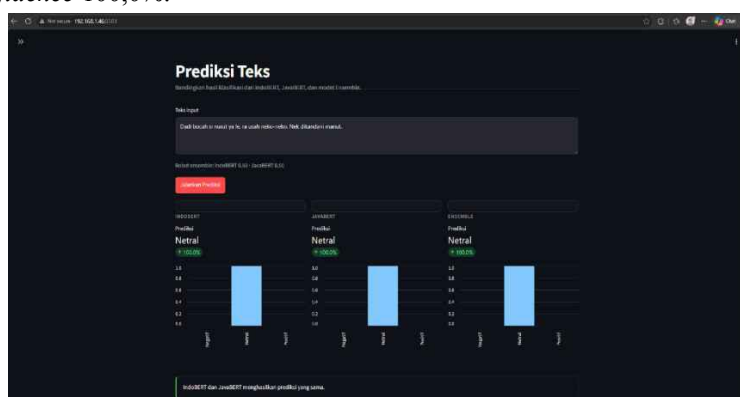
Gambar 11. Prediksi Text Pada Kasus A

Pada **Kasus A**, kalimat berdialek Ngapak “*Inyong kue bisa garap apa bae, jajal di test bae men terbukti*” (kira-kira: “Saya bisa mengerjakan apa saja, coba saja di-*test* agar terbukti”) menghasilkan prediksi yang **berbeda** antar-model: IndoBERT memprediksi Netral (100,0%), JavaBERT memprediksi Negatif (100,0%), sedangkan model *ensemble* hasil rata-rata kedua probabilitas yang saling bertentangan berakhir pada kondisi hampir seri dengan Negatif hanya unggul tipis (50,0% berbanding 50,0% pada Netral). Kalimat ini menggunakan kosakata khas Ngapak seperti *inyong* (aku/saya), *kue* (kowe/kamu), *bae* (wae/saja), dan *men* (supaya) yang tidak terdapat dalam leksikon Jawa standar yang digunakan untuk pelabelan maupun lazim ditemukan pada korpus latih, sehingga kedua model pada dasarnya menghadapi kosakata di luar distribusi (*out-of-vocabulary/out-of-distribution*).



Gambar 12. Prediksi Text Pada Kasus B

Pada **Kasus B**, kalimat “*Dadi bocah angel temen, di kandani ora mudheng-mudheng*” (Jadi anak susah sekali, dikasih tahu tidak mengerti-mengerti) menghasilkan kesepakatan penuh ketiga model pada label Negatif dengan *confidence* 100,0%.



Gambar 13. Prediksi Text Pada Kasus C

Pada **Kasus C**, kalimat “*Dadi bocah sing nurut ya le, ra usah neko-neko. Nek dikandani manut.*” (Jadi anak yang menurut ya nak, tidak usah macam-macam. Kalau dikasih tahu, patuh.) menghasilkan kesepakatan penuh pada label Netral dengan *confidence* 100,0%. Kedua kalimat ini menggunakan Bahasa Jawa *ngoko* standar dengan penanda leksikal yang relatif eksplisit dan sejalan dengan pola leksikon pelabelan.

3.4 Interpretasi Skor Evaluasi Sempurna

Skor akurasi, presisi, *recall*, dan F1 sebesar 1,00 pada seluruh model dan seluruh kelas (Tabel 4-5) secara statistik tidak lazim untuk tugas analisis sentimen berbasis bahasa alami, dan perlu dibaca sebagai artefak metodologis, bukan sebagai indikasi kemampuan generalisasi sempurna. Penyebab utamanya adalah label pada seluruh data (latih, validasi, dan uji) dibangkitkan oleh aturan leksikon yang sama dan bersifat deterministik terhadap kemunculan kata kunci tertentu. Karena kata-kata kunci tersebut secara eksplisit muncul sebagai substring pada teks yang dilabelinya, hubungan antara fitur permukaan (kata) dan label menjadi sangat kuat dan hampir linier-terpisah (*near-linearly separable*) sebuah pola yang mudah dipelajari secara hampir sempurna oleh model kapasitas tinggi seperti BERT, sebagaimana lazim terjadi pada fenomena shortcut learning [10] dan heuristik permukaan (*surface heuristics*) yang telah didokumentasikan pada berbagai benchmark NLP lain [11], [12], setelah tiga *epoch fine-tuning*, sebagaimana tercermin dari *training loss* yang sangat rendah (0,014-0,022). Dengan kata lain, ketiga model kemungkinan besar tidak mempelajari pemahaman sentimen tingkat pragmatik, melainkan memetakan kembali (*re-learning*) sinyal leksikon yang sama yang digunakan untuk membangkitkan label ground-truth itu sendiri. Split data *train/validation/test* secara acak tidak dapat memitigasi isu ini karena kebocoran informasi (*label leakage*) terjadi pada tingkat skema pelabelan, bukan pada tingkat pembagian data.

3.5 Bukti Pendukung dari Pengujian Kualitatif

Temuan pada Bagian 3.2 dan 3.3 memperkuat interpretasi di atas sekaligus memberikan gambaran performa yang lebih realistis. Pada kalimat dialek standar dengan penanda leksikal eksplisit (Kasus B dan C, serta empat dari lima kalimat pada Tabel 6), ketiga model konsisten dan sangat percaya diri sejalan dengan hipotesis bahwa model telah mempelajari sinyal leksikon dengan sangat baik. Namun pada kalimat yang menyimpang dari pola leksikal yang dipelajari baik karena dialek non-standar (Kasus A: Ngapak) maupun karena register informal/retoris (kalimat makian pada Bagian 3.2) performa model menurun signifikan, ditandai dengan perbedaan pendapat antar-model dan *confidence ensemble* yang mendekati 50:50. Kasus “*Regane biasa wae, ora ono sing istimewa*” yang diklasifikasikan Negatif alih-alih Netral turut menunjukkan

bahwa model condong mengasosiasikan frasa negasi apa pun dengan sentimen negatif, alih-alih membedakan negasi lemah (netral) dari negasi kuat (negatif) secara pragmatis.

3.6 Perbandingan IndoBERT, JavaBERT, dan Ensemble

Pada metrik evaluasi standar, ketiga model tidak dapat dibedakan (Tabel 5) karena keterbatasan metodologis di atas. Akan tetapi, studi kasus Ngapak (Kasus A) memberikan sinyal pembeda yang berharga: IndoBERT dan JavaBERT menghasilkan prediksi yang berlawanan, mengindikasikan bahwa kedua model memiliki pola generalisasi kosakata di luar distribusi yang berbeda meskipun performanya identik pada data uji standar. Dalam kondisi disagreement seperti ini, skema *ensemble soft-voting* dengan bobot setara ($w_I = w_J = 0,5$) tidak serta-merta menghasilkan keputusan yang lebih baik, melainkan hanya merata-ratakan ketidaksepakatan menjadi keputusan yang nyaris acak (mendekati 50:50). Hal ini menunjukkan bahwa *ensemble* berbasis pembobotan F1 validasi yang pada penelitian ini bernilai identik (1,0000) untuk kedua model karena isu *label leakage* yang sama bukan skema pembobotan yang optimal untuk kondisi di luar distribusi, dan skema *ensemble* yang lebih adaptif (misalnya berbasis *confidence* per-sampel atau *stacking*) berpotensi lebih sesuai untuk kasus semacam ini.

4. KETERBATASAN DAN ARAH PENELITIAN SELANJUTNYA

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Seluruh label ground-truth dibangkitkan secara otomatis melalui aturan leksikon, sehingga model pada dasarnya mempelajari kembali sinyal leksikon yang sama yang digunakan untuk membangkitkan labelnya sendiri (label leakage), bukan pemahaman sentimen pada tingkat pragmatik sebagaimana dibahas pada Bagian 3.4, ini menjelaskan mengapa skor evaluasi mendekati sempurna dan tidak dapat diatasi hanya dengan pembagian data yang berbeda. Ke depan, hal ini perlu diatasi dengan membangun benchmark human-annotated yang independen dari skema leksikon, melibatkan multiple annotator dan pengukuran inter-annotator agreement [13]; active learning dapat diterapkan untuk memprioritaskan anotasi manusia pada sampel paling ambigu [14].

Cakupan dialek dalam penelitian ini juga masih terbatas, sebagaimana ditunjukkan secara konkret pada Kasus A: kalimat berdialek Ngapak dengan kosakata seperti *inyong*, *kue*, *bae*, dan *men* menghasilkan prediksi yang tidak konsisten antar-model (IndoBERT Netral, JavaBERT Negatif) hingga model ensemble berakhir pada kondisi hampir seri, berbeda dengan Kasus B dan C yang menggunakan Jawa ngoko standar dan menghasilkan kesepakatan penuh pada ketiga model. Hal ini mengindikasikan bahwa leksikon dan data latih tidak cukup merepresentasikan dialek regional seperti Ngapak, Osing, maupun varian Jawa Suriname, sebuah pola kesenjangan performa lintas-dialek yang juga telah didokumentasikan pada bahasa lain [21]; riset selanjutnya perlu memperluas cakupan ini dengan menyusun test set out-of-dialect secara terpisah. Selain itu, dataset berasal dari satu sumber daring (Kaggle) dengan gaya penulisan relatif homogen, sehingga generalisasinya ke domain lain (media sosial, berita) belum teruji.

Evaluasi juga hanya menggunakan satu kali pembagian train/validation/test tanpa cross-validation, dan pengujian kualitatif di luar distribusi baru mencakup delapan kalimat jumlah yang belum cukup untuk klaim generalisasi kuantitatif yang kuat. Ke depan, skema ensemble yang lebih adaptif seperti stacking dengan meta-learner, pembobotan berbasis confidence, atau mekanisme abstain untuk kasus low-confidence seperti pada Kasus A perlu dieksplorasi, sekaligus mengoptimalkan model untuk deployment ringan (distilasi/kuantisasi) [15], [20] mengingat potensi aplikasi real-time pada antarmuka interaktif yang telah dibangun (Gambar 11).

5. KESIMPULAN

Penelitian ini menyajikan studi komparatif IndoBERT, JavaBERT, dan model *ensemble soft-voting* keduanya untuk analisis sentimen multikelas Bahasa Jawa, dilengkapi dengan skema pelabelan otomatis berbasis leksikon untuk mengatasi ketiadaan data berlabel. Ketiga model mencapai skor evaluasi sempurna (akurasi/F1 = 1,00) pada data uji standar, namun penelitian ini menunjukkan secara kritis bahwa hasil tersebut merupakan konsekuensi dari kuatnya korelasi antara skema pelabelan berbasis leksikon dan fitur leksikal yang dipelajari model, bukan indikasi pemahaman sentimen yang mendalam. Pengujian kualitatif tambahan pada kalimat berdialek Ngapak mengungkap keterbatasan generalisasi yang nyata: IndoBERT dan JavaBERT saling berbeda pendapat, dan model *ensemble* berakhir pada keputusan yang hampir seri. Temuan ini menggarisbawahi pentingnya evaluasi di luar distribusi dan benchmark berlabel manusia dalam menilai kesiapan model analisis sentimen Bahasa Jawa untuk aplikasi dunia nyata, serta membuka arah penelitian lanjutan yang jelas seputar cakupan dialek, skema *ensemble* adaptif, dan validasi anotasi manusia.

DAFTAR PUSTAKA

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171-4186.
- [2] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics and 10th Int. Joint Conf. Natural Language Processing (AACL-IJCNLP)*, Suzhou, China, 2020, pp. 843-857.
- [3] W. Wongso, D. S. Setiawan, and D. Suhartono, "Causal and masked language modeling of Javanese language using transformer-based architectures," in *Proc. 2021 Int. Conf. Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia, 2021, pp. 1-7.
- [4] G. I. Winata et al., "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," in *Proc. 17th Conf. European Chapter Assoc. Comput. Linguistics (EACL)*, Dubrovnik, Croatia, 2023, pp. 815-834.
- [5] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093-1113, 2014.
- [6] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. 1st Int. Workshop on Multiple Classifier Systems (MCS)*, Cagliari, Italy, 2000, pp. 1-15.
- [7] Kaggle, "Javanese Sentiment Dataset." [Online]. Tersedia: <https://www.kaggle.com/datasets/nunoliebert/javanese-sentiment-dataset>. [Diakses: 5 Agustus 2026].
- [8] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1-135, 2008.
- [9] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguist.*, vol. 37, no. 2, pp. 267-307, 2011.
- [10] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665-673, 2020.
- [11] S. Gururangan et al., "Annotation artifacts in natural language inference data," in *Proc. 2018 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), Vol. 2 (Short Papers)*, New Orleans, LA, USA, 2018, pp. 107-112.
- [12] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Florence, Italy, 2019, pp. 3428-3448.
- [13] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159-174, 1977.
- [14] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, *University of Wisconsin-Madison*, 2009.
- [15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [16] F. Koto et al., "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. 28th Int. Conf. Comput. Linguistics (COLING)*, Barcelona, Spain, 2020, pp. 757-770.
- [17] A. F. Aji et al., "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 7226-7249.
- [18] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998-6008.
- [19] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Proc. 18th China National Conf. Chinese Computational Linguistics (CCL)*, Kunming, China, 2019, pp. 194-206.
- [20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [21] S. L. Blodgett, L. Green, and B. O'Connor, "Demographic dialectal variation in social media: A case study of African-American English," in *Proc. 2016 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Austin, TX, USA, 2016, pp. 1119-1130.