Part-of-Speech Tagging Bahasa Jawa Menggunakan Model Pre-Trained Bidirectional Encoder Representation from Transformers

Ahmad Izzuddin[1], Nuzul Hikmah[2], Muhammad Alvin Ajry*[3]
Program Studi Teknik Elektro Fakultas Teknik dan Informatika Universitas Panca Marga
Jl. Raya Dringu, Krajan, Pabean, Kec. Mayangan, Probolinggo, 67212
e-mail: [1]ahmad.izzuddin@upm.ac.id, [2]n.hikmah1807@upm.ac.id, [3]alfinazry791@gmail.com
*Penulis Korespondensi

Diterima: 14 Oktober 2025; Direvisi: 12 Mei 2026; Disetujui: 15 Mei 2026

Abstrak

Part-of-Speech Tagging (POS tagging) merupakan proses penentuan kelas kata dalam suatu teks yang penting dalam pemrosesan bahasa alami (Natural Language Processing). Pada bahasa Jawa, POS tagging masih merupakan tantangan karena keterbatasan sumber daya linguistik dan kompleksitas bahasa tersebut. Dengan perkembangan teknologi deep learning, metode fine-tuning BERT (Bidirectional Encoder Representations from Transformers) telah diterapkan untuk melakukan penandaan kelas kata dalam bahasa Jawa, yang merupakan bahasa dengan sumber daya terbatas. Model javanese-bert-small dilatih menggunakan dataset UD_Javanese-CSUI, dan dievaluasi menggunakan metrik precision, recall, F1-score, dan accuracy. Hasil penelitian menunjukkan bahwa model mencapai performa mumpuni dengan akurasi tercapai 88,87%, serta menunjukkan kestabilan selama pelatihan tanpa overfitting signifikan. Temuan ini menunjukkan bahwa pendekatan berbasis BERT efektif untuk menangani ambiguitas kelas kata dalam bahasa Jawa dan dapat menjadi pijakan untuk pengembangan lebih lanjut dalam sistem NLP untuk bahasa daerah.

Kata kunci: *bahasa jawa, deep learning, bert, part-of-speech tagging*

Abstract

Part-of-Speech Tagging (POS tagging) is the process of determining word classes in a text that is important in natural language processing. In Javanese, POS tagging is still a challenge due to limited linguistic resources and the complexity of the language. With the development of deep learning technology, the BERT (Bidirectional Encoder Representations from Transformers) fine-tuning method has been applied to classify word classes in Javanese, which is a language with limited resources. The javanese-bert-small model was trained using the UD_Javanese-CSUI dataset, and evaluated using precision, recall, F1-score, and accuracy metrics. The results showed that the model achieved good performance with an accuracy of 88,87%, and showed stability during training without significant overfitting. These findings indicate that the BERT-based approach is effective in handling word class ambiguity in Javanese and can be a stepping stone for further development in NLP systems for regional languages.

Keywords*: bert, deep learning, javanese language, part-of-speech tagging*

I. PENDAHULUAN

Generasi muda di wilayah Jawa mengalami penurunan signifikan dalam kemahiran berbahasa Jawa. Data menunjukkan bahwa hanya 61,7% anak-anak di Jawa yang masih aktif menggunakan bahasa Jawa dalam percakapan sehari-hari, dan tren ini terus menurun akibat pengaruh globalisasi dan preferensi bahasa Indonesia sebagai bahasa komunikasi utama [1]. Tren ini membahayakan pelestarian tradisi linguistik Jawa, yang merupakan warisan budaya bagi lebih dari 68 juta penuturnya [2].

Inovasi dalam metode pembelajaran bahasa Jawa melalui teknologi digital sangat diperlukan, khususnya aplikasi pembelajaran bahasa interaktif [3], [4], [5]. Platform pembelajaran yang sukses membutuhkan sistem komputasi canggih yang mampu memahami dan menganalisis pola linguistik dengan

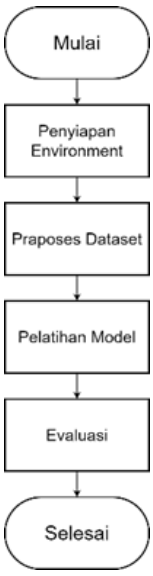
tepat melalui metodologi *Natural Language Processing* (NLP) terkini [6], [7], [8], [9], [10]. Tantangan mendasar adalah kompleksitas bahasa Jawa yang memiliki ambiguitas leksikal tinggi, di mana satu kata dapat memiliki fungsi atau makna berbeda tergantung konteks [11], [12]. Sebagai contoh, dalam kalimat "*Tuku beras iku ora tuku murah*" (artinya: "Membeli beras itu tidak berarti membeli dengan harga murah"), kata "*tuku*" yang pertama berfungsi sebagai kata kerja ("membeli"), sementara kata "*tuku*" yang kedua berfungsi sebagai kata benda (merujuk pada "pembelian"). Ketidakakuratan dalam menafsirkan ambiguitas ini akan memengaruhi sistem pembelajaran dalam memberikan koreksi yang tepat [11], [12], [13]. Dengan demikian, mekanisme *Part-of-Speech tagging* yang presisi untuk bahasa Jawa menjadi prasyarat teknis penting untuk membangun platform pendidikan bahasa yang sukses dan responsif terhadap keragaman konteks penggunaan bahasa Jawa.

Berbagai pendekatan telah dikembangkan untuk menangani tugas POS *tagging* pada bahasa Jawa. Penelitian menggunakan *Conditional Random Fields* (CRF) mencapai akurasi 67% pada dataset 3.000 kata bahasa Jawa Krama [14], namun metode ini kesulitan menangkap hubungan kontekstual jarak jauh dan kompleksitas morfologi bahasa Jawa [15], [16], [17]. Pendekatan *Hidden Markov Model* (HMM) menunjukkan performa lebih tinggi dengan akurasi 92,6% pada dataset 1.770 kata [18], namun hasil ini didapat dari evaluasi dengan dataset pengujian yang terlalu mirip dengan pelatihan, mengakibatkan *overfitting* dan kurangnya penilaian otentik terhadap kemampuan generalisasi [18]. Penelitian terbaru memanfaatkan perkembangan *deep learning* dengan model berbasis *Transformer*. Penerapan *fine-tuning* BERT pada ambiguitas leksikal dalam POS *tagging* mencapai *F1-Score* 0,9656 [19], [20], [21], menunjukkan efektivitas pendekatan *transfer learning*. Namun, penelitian ini menggunakan model *multilingual* BERT (lebih berat) dan memerlukan sumber daya komputasi tinggi, serta terkadang gagal menangkap konteks kalimat tertentu [19], [20], [21].

Penelitian ini memiliki kebaruan berupa penerapan pendekatan *transfer learning* dengan model *javanese-bert-small* yang di-*fine-tune* menggunakan korpus UD_Javanese-CSUI berkualitas tinggi untuk menangani ambiguitas leksikal tingkat tinggi dalam POS *tagging* bahasa Jawa. Berbeda dengan penelitian sebelumnya yang menggunakan CRF atau HMM dengan dataset terbatas, pendekatan ini memanfaatkan *pretrained transformer* yang telah dilatih pada corpus Indonesia untuk memahami pola linguistik Jawa. Tujuan penelitian ini adalah mengembangkan sistem POS *tagging* untuk bahasa Jawa yang mencapai akurasi kompetitif (target 90%) sambil memastikan generalisasi yang kuat dan efisiensi komputasi

II. METODE PENELITIAN

Penelitian ini menggunakan *Google Colab* dimanfaatkan sebagai platform untuk melatih model. Berikut pada gamabr 1 merupakan alur penelitiannya:



Gambar 1. Tahapan penelitian

A. Penyiapan *Environment*

Tahapan ini meliputi pengaturan lingkungan penelitian, mulai dari inisialisasi *Google Colab* hingga pemasangan pustaka-pustaka pendukung terutama conllu, torch, seqeval, dan evaluate.

B. Praproses Dataset

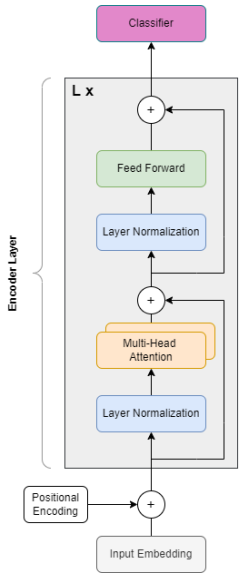
Tahapan kedua dimulai dengan data mentah yang diolah dengan mengidentifikasi jumlah kalimat dan label, lalu diubah ke format yang sesuai untuk model. Selanjutnya, kalimat dipecah menjadi token dan dicocokkan dengan label *tag*-nya. Dataset yang digunakan berasal dari *UD-Javanese-CSUI* yang bersumber dari *Universal Dependencies* yang dapat diakses melalui tautan berikut (https://universaldependencies.org/treebanks/jv_csui/index.html#ud-javanese-csui). Dataset ini berisi 1.000 kalimat dengan 14.000 kata dalam format CoNLL-U, serta mencakup 17 kelas kata [2]. Pembagian data mengikuti praktik standar *machine learning*, yaitu 80% untuk pelatihan (800 sampel), 10% untuk validasi (100 sampel), dan 10% untuk pengujian (100 sampel). Kelas kata yang digunakan dapat dilihat pada tabel 1 berikut:

Tabel 1. Rincian kelas kata

Kelas Kata	Singkatan	Contoh
Kata Sifat	ADJ	*apik* (bagus), *abot* (berat), *fast* (cepat)
Adposisi	ADP	*saka* (dari), *marang* (ke), *ing* (di)
Kata Keterangan	ADV	*mung* (hanya), *wis* (sudah), *isih* (masih)
Kata Bantu	AUX	*arep* (akan), *kudu* (harus), *bakal* (akan)
Konjungsi Koordinatif	CCONJ	*lan* (dan), *utawa* (atau), *nanging* (tapi)
Penentu	DET	*iki* (ini), *kuwi* (itu), *siji* (satu)
Interjeksi	INTJ	*lho* (tahu kan), *wah*, eh
Kata Benda	NOUN	*anak* (anak), *omah* (rumah), *pitik* (ayam)
Numeralia	NUM	*loro* (dua), *telu* (tiga), *ten* (sepuluh)
Partikel	PART	*wae* (saja), *really, anyway*
Pronomina	PRON	*aku* (saya), *kowe* (kamu), *dheweke* (dia)
Nama Diri	PROPN	Surabaya, Jokowi, Indonesia
Tanda Baca	PUNCT	. , ? !
Konjungsi Subordinatif	SCONJ	*amarga* (karena), *nek* (jika), agar (agar)
Simbol	SYM	%, +, =
Kata Kerja	VERB	*manganese* (makan), *mlaku* (berjalan), *ngombe* (minum)
Lain-lain	X	*e-mail, download*

C. Pelatihan Model

Tahapan selanjutnya melakukan pelatihan pada model dengan proses dari membuat prediksi, menghitung kesalahan, dan memperbaiki diri secara berulang selama model dilatih.



Gambar 2. Arsitekur bert

Setelah setiap perulangan, model diuji dengan data validasi untuk mengukur kinerjanya. Penelitian ini menggunakan *model javanese-bert-small* dari *Hugging Face Hub* (https://huggingface.co/w11wo/javanese-bert-small) (Gambar 2). Model ini dirancang khusus untuk pemrosesan bahasa Jawa [22].

D. Evaluasi

Tahapan terakhir yaitu mengevaluasi kinerja model menggunakan metrik *precision, recall, f1-score,* dan *accuracy*. Selanjutnya evaluasi dianalisis untuk mengidentifikasi pola kesalahan yang sering terjadi. Empat metrik memiliki persamaan yang ditunjukkan dalam persamaan (1)-(4)_sebagai berikut.

1) *Precision*: Metrik yang mengukur keakuratan prediksi positif dengan menghitung proporsi kasus positif sejati dari seluruh kasus yang diprediksi sebagai positif. Ini memastikan model meminimalkan kesalahan penandaan data negatif menjadi positif.

$$Precision = TP/(TP + FP) \quad (1)$$

2) *Recall*: Metrik yang mengukur proporsi kasus positif aktual yang berhasil diidentifikasi dengan benar oleh model. Metrik ini sangat penting untuk meminimalkan kasus positif yang ditandakan sebagai negatif.

$$Recall = TP/(TP + FN) \quad (2)$$

3) *F1-Score*: Metrik yang megukur rata-rata harmonis dari *precision* dan *recall* yang menawarkan keseimbangan antara keduanya. Metrik ini sangat efektif untuk digunakan pada kumpulan data yang tidak seimbang karena memperhitungkan baik *false positive* maupun *false negative*.

$$F1 - Score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (3)$$

4) *Accuracy*: Metrik yang mengukur seberapa sering model membuat prediksi yang benar secara keseluruhan. Ini adalah rasio prediksi yang benar, baik positif maupun negatif terhadap total data.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

III. HASIL DAN PEMBAHASAN

A. Pelatihan dan Evaluasi Model

Hasil pelatihan model menunjukkan konvergensi pembelajaran yang efektif selama 35 *epoch*, yang menjamin hasil eksperimen dapat diulang. Model mencapai konvergensi optimal dengan loss pelatihan akhir sebesar 0,316, menandakan keberhasilan dalam mempelajari pola linguistik bahasa Jawa. Pola konvergensi ini terlihat jelas dari epoch 1 hingga epoch 10, di mana terjadi penurunan loss yang signifikan dari 2,5 menjadi sekitar 0,1, diikuti oleh stabilisasi pada epoch berikutnya. Data hasil pelatihan model terdapat pada tabel 2 berikut ini.

Tabel 2. Hasil pelatihan model

Epoch	*Training Loss*	*Validation loss*	*Precission*	*Recal*	*F1-Score*	*Accuracy*
1	-	2.534	0.253	0.093	0.136	0.267
10	0.162	0.418	0.852	0.852	0.852	0.881
20	0.018	0.539	0.862	0.866	0.864	0.889
30	0.007	0.571	0.869	0.868	0.869	0.894
35	0.005	0.580	0.868	0.866	0.867	0.892

Nilai kerugian validasi yang relatif rendah (0,571), jika dibandingkan dengan kompleksitas tugas penandaan, membuktikan bahwa model berhasil mempelajari fitur diskriminatif untuk setiap kategori kelas kata bahasa Jawa. Kinerja yang konsisten pada data validasi, dengan pola yang serupa dengan saat pelatihan, menunjukkan bahwa model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang kuat. Data performa validasi dicantumkan pada tabel 3 berikut ini.

Tabel 3. Performa pada data validasi

Metrik	Statistik
accuracy	0.8943710910354413
f1-score	0.8691222570532915
precision	0.8698039215686274
recall	0.8684416601409554
loss (kerugian)	0.5713428258895874
runtime (waktu jalan)	0.3521
samples per second (sampel per detik)	284.014
steps per second (langkah per detik)	19.881

Selama evaluasi pengujian dalam tabel 4, sistem menunjukkan efektivitas yang stabil dengan mencapai akurasi 88,87%, *F1-score* 86,33%, presisi 86,59%, dan *recall* 86,07%, disertai nilai loss sebesar 0,643. Hasil *F1-score* menunjukkan keseimbangan ideal antara metrik presisi dan *recall*. Ini membuktikan bahwa sistem dapat mempertahankan akurasi dalam penandaan positif sambil berhasil meminimalkan kasus *false negative*. Hasil presisi yang tinggi menunjukkan akurasi model dalam memprediksi positif dan kemampuannya untuk meminimalkan kesalahan penandaan *false positive*. Sementara itu, hasil recall menunjukkan kemampuan model yang baik dalam mendeteksi sebagian besar kasus yang seharusnya ditandai ke dalam kategori tertentu. Nilai kerugian yang tercatat berada dalam rentang yang dapat diterima untuk tugas penandaan, menunjukkan prediksi model yang meyakinkan. Perbedaan kinerja antara data validasi dan data uji juga sangat minim, dengan perbedaan akurasi hanya 0,57% (89,44% vs 88,87%), perbedaan *F1-score* 0,58% (86,91% vs 86,33%), dan metrik lainnya yang juga menunjukkan konsistensi tinggi.

Tabel 4. Performa pada data uji

Metrik	Statistik
accuracy	0.8887408394403731
f1-score	0.863260706235913
precision	0.86586284853052
recall	0.8606741573033708
loss (kerugian)	0.6431677937507629
runtime (waktu jalan)	0.2619
samples per second (sampel per detik)	284.014
steps per second (langkah per detik)	19.881

B. Performa Kelas Kata

Berdasarkan tabel 5 hasil dari performa model, beberapa kelas kata mencapai kinerja tinggi dengan *F1-score* 0,90 atau lebih:

1) PUNCT (Tanda Baca): Kelas yang meraih hasil baik dengan presisi, *recall*, dan *F1-score* 1,00. Hal ini disebabkan oleh sifat simbol tanda baca yang seragam dan mudah dikenali seperti titik, koma, atau tanda tanya.
2) AUX (Kata Bantu): Kelas yang menunjukkan efektivitas yang hampir sempurna dengan presisi 0,97 dan *recall* 1,00, menghasilkan *F1-score* 0,99. Ini membuktikan kemampuan sistem yang luar biasa dalam mengidentifikasi kata-kata seperti "*bakal*", "*arep*", atau "*wis*" yang berfungsi mendukung kata kerja utama.
3) PART (Partikel): Kelas yang memperoleh presisi 1,00 dan *recall* 0,92, dengan *F1-score* 0,96. Kategori ini mencakup kata-kata pendek seperti "*kok*", "*lho*", atau "*ta*" yang memberikan makna khusus pada struktur kalimat bahasa Jawa.
4) CCONJ (Konjungsi Koordinatif): Kelas yang menunjukkan kinerja baik dengan presisi sempurna (1,00) dan *recall* 0,94, mencapai *F1-score* 0,97. Kata-kata seperti "*lan*", "*utawa*", atau "*nanging*" yang berfungsi sebagai penghubung antar komponen kalimat berhasil diidentifikasi dengan sangat akurat karena pola penggunaannya yang stabil.

Hasil moderat juga dihasilkan oleh model, dengan nilai F1-score antara 0,80 hingga 0,90:

1) NOUN (Kata Benda): Kelas sebagai kategori dengan data terbanyak (296 instans), menunjukkan hasil yang baik dengan presisi 0,86, *recall* 0,92, dan *F1-score* 0,89. Hasil ini dapat diterima mengingat kata

benda bahasa Jawa memiliki variasi yang luas dan dapat mengalami perubahan morfologi dengan penambahan awalan atau akhiran.

2) PROPN (Nama Diri): Kelas yang mencakup 167 data, menunjukkan keseimbangan yang baik antara akurasi (0,89) dan sensitivitas (0,88), menghasilkan *F1-score* 0,89. Sistem menunjukkan kemampuan kuat untuk membedakan kata benda biasa seperti "*omah*" dari nama diri seperti "Jakarta" atau "Slamet".
3) VERB (Kata Kerja): Kelas dengan 192 instans, memperoleh *F1-score* 0,87, dengan presisi 0,89 dan *recall* 0,84. Temuan ini sangat baik mengingat kata kerja bahasa Jawa dapat mengalami berbagai perubahan morfologi, seperti penambahan awalan "*ng*-" atau akhiran "*-ake*" dan "*-i*".
4) ADP (Adposisi): Kelas yang menunjukkan efektivitas yang stabil dengan *F1-score* 0,94, *presisi* 0,95, dan *recall* 0,93. Kata-kata seperti "*ing*", "*saka*", atau "*menyang*" yang menunjukkan hubungan spasial, temporal, atau direksional berhasil diidentifikasi dengan baik karena pola penggunaannya yang jelas.

Model juga menunjukkan hasil yang cukup memadai, tercermin dari *F1-score* yang relatif rendah dan membutuhkan penanganan untuk pengembangan selanjutnya:

1) ADJ (Kata Sifat): menunjukkan hasil terlemah dengan *F1-score* 0,76, presisi 0,74, dan *recall* 0,77. Kinerja yang rendah ini berasal dari kecenderungan kata sifat bahasa Jawa untuk berfungsi ganda sebagai kata deskriptif maupun kata benda tergantung pada konteksnya. Sebagai contoh, "apik" bisa berarti "bagus" (kata sifat) atau merujuk pada "kebaikan" (kata benda).
2) INTJ (Interjeksi): menunjukkan presisi sempurna (1,00) tetapi mengalami kinerja *recall* yang buruk (0,60), menghasilkan *F1-score* 0,75. Hal ini diakibatkan oleh contoh data pelatihan yang tidak memadai (hanya 5 instans), sehingga model sulit membentuk pola identifikasi yang andal untuk kata seru seperti "*aduh*", "*wah*", atau "*lho*".
3) SCONJ (Konjungsi Subordinatif): menunjukkan *F1-score* 0,82, dengan presisi 0,77 dan sensitivitas 0,88. Kata-kata seperti "*amarga*", "*nalika*", dan "*supaya*" berfungsi menghubungkan kalimat utama dengan frasa penjelas. Namun, konektor ini sulit diidentifikasi karena sering tumpang tindih secara fungsional dengan kategori tata bahasa lain dalam berbagai konteks kalimat.

Tabel 5. Hasil performa per kelas kata

Kelas Kata	*Precision*	*Recall*	*F1*	*Support*
ADJ	0.7397	0.7714	0.7552	70
ADP	0.9512	0.9286	0.9398	84
CCONJ	1.0000	0.9412	0.9697	34
INTJ	1.0000	0.6000	0.7500	5
NUM	0.8947	0.9444	0.9189	18
PRON	0.9059	0.8191	0.8603	94
PUNCT	1.0000	1.0000	1.0000	192
VERB	0.8944	0.8385	0.8656	192
ADV	0.8488	0.7935	0.8202	92
AUX	0.9744	1.0000	0.9870	38
DET	0.7907	0.9067	0.8447	75
NOUN	0.8631	0.9155	0.8885	296
PART	1.0000	0.9231	0.9600	26
PROPN	0.8909	0.8802	0.8855	167
SCONJ	0.7692	0.8824	0.8219	34
X	0.8415	0.8214	0.8313	84
macro avg	**0.8978**	**0.8729**	**0.8812**	**1501**
weighted avg	**0.8909**	**0.8887**	**0.8888**	**1501**
accuracy			**0.8887**	**1501**

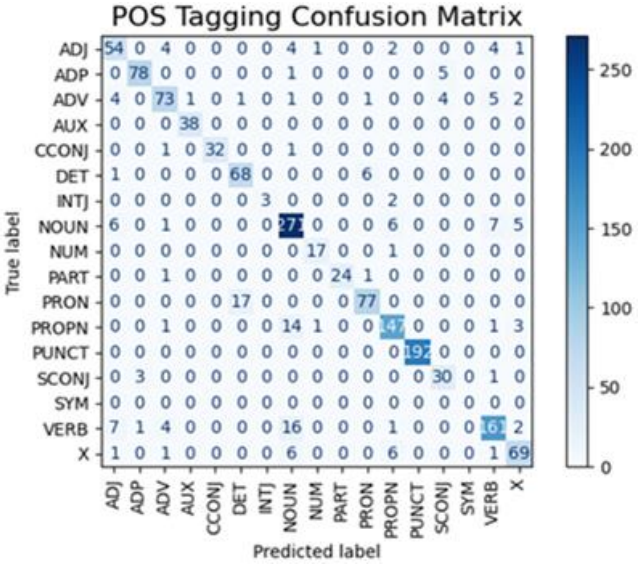
Berdasarkan matriks kebingungan pada Gambar 3, matriks diagonal menunjukkan penandaan yang benar. Angka-angka berwarna gelap pada diagonal, seperti 54 untuk ADJ (kata sifat), 78 untuk ADP (adposisi), dan 73 untuk ADV (kata keterangan), menunjukkan jumlah prediksi yang akurat untuk setiap kategori. Pola kesalahan penandaan yang paling menonjol terlihat pada beberapa hal:

1) Kategori ADJ (kata sifat) mengalami kesalahan terbesar, di mana 7 kata sifat salah diidentifikasi sebagai NOUN (kata benda) dan 4 kata sifat lainnya salah ditandai sebagai ADV (kata keterangan). Pola kesalahan

ini mencerminkan sifat bahasa Jawa yang memungkinkan satu kata memiliki fungsi ganda tergantung pada posisi dan konteks penggunaannya dalam kalimat.

2) Kesalahan lain terjadi pada kategori VERB (kata kerja), di mana matriks menunjukkan beberapa kasus salah penandaan sebagai NOUN. Hal ini disebabkan oleh proses morfologi bahasa Jawa yang memungkinkan kata kerja berubah fungsi menjadi kata benda melalui nominalisasi dengan penambahan imbuhan tertentu.

3) Kategori DET (penentu) juga menunjukkan pola kesalahan yang menarik, di mana beberapa kasus salah ditandai sebagai PRON (pronomina). Matriks menunjukkan kebingungan ini melalui angka di luar diagonal utama pada baris DET dan kolom PRON. Model mengalami kesulitan membedakan kata seperti "ki" yang dapat berfungsi sebagai penentu ataupun sebagai pronomina, tergantung pada konteks kalimat.



Gambar 3. *Confusion matrix*

C. Hasil Prediksi dan Analisis Kesalahan

Berdasarkan tabel 6 hasil prediksi pada kalimat, berikut rincian mengenai hasil prediksi yang diperoleh dari skenario pengujian ini:

1) Kalimat dengan Struktur Sederhana: Hasil prediksi pada kalimat pertama "*Ing njeron gang laladan Kampung Batik Kauman ana saperangan signboard werna putih kang pasang ing cagak werna ijo tuwa*" menunjukkan akurasi sempurna 100%. Model berhasil mengidentifikasi setiap elemen kalimat dengan benar, mulai dari preposisi "*Ing*", kata benda, nama tempat "Kampung Batik Kauman", kata kerja, penentu, kata sifat, hingga pronomina relatif. Temuan ini membuktikan bahwa sistem efektif dalam memahami struktur sintaksis bahasa Jawa yang rumit.

2) Kalimat dengan Kompleksitas Sedang: Hasil tes pada kalimat kedua mencapai akurasi 91,3%. Kesalahan pertama terjadi pada kata "*punkasan*" yang diprediksi sebagai NOUN padahal kelas kata aslinya ADJ, menunjukkan ambiguitas fungsi kata. Kesalahan kedua adalah kegagalan prediksi pada kata "*pamekaran*" yang diberi simbol X, menandakan keterbatasan model dalam mengenali bentuk derivasional kompleks dengan imbuhan "*pa-*" dan "*-an*" yang merupakan ciri khas morfologi bahasa Jawa. Pada kalimat ketiga, akurasi mencapai 80% dengan pola kesalahan yang lebih beragam. Pola ini mencakup kesulitan dalam membedakan kata sifat dan kata kerja statis, serta generalisasi terhadap urutan kata yang secara struktural menyerupai nama diri.

3) Kalimat dengan Tingkat Kesulitan Tinggi: Hasil tes prediksi pada kalimat keempat mewakili kasus terendah dengan akurasi 73,3%, mengungkapkan beberapa keterbatasan fundamental sistem. Kesalahan morfologi kompleks, terjadi pada kata "*Jarene*" yang sama sekali tidak dapat diprediksi. Kata ini merupakan bentuk evidential yang menyampaikan informasi tidak langsung, sebuah kategori gramatikal yang membutuhkan pemahaman kontekstual yang mendalam. Kesalahan kategori gramatikal, model salah menandakan beberapa kata, seperti "*donga*" yang diprediksi sebagai NOUN (seharusnya VERB), dan "*nesu*" yang diprediksi sebagai VERB (seharusnya ADJ). Kesalahan struktural, model menunjukkan

keterbatasan dalam memahami fungsi sintaksis multi-kategori, termasuk kebingungan antara fungsi preposisional dan konjungsi dari kata yang sama tergantung pada konteks strukturalnya.

Tabel 6. Hasil prediksi pada kalimat

Kalimat	Kategori	Total Token	Token Benar	Akurasi	Kesalahan
1	Berhasil	20	20	100.0%	-
2	Menengah	23	21	91.3%	*pamekaran* (X), *pungkasan* (ADJ→NOUN)
3	Menengah	25	20	80.0%	*puntir* (VERB→NOUN), *kapisah* (VERB→ADJ), field of attraction (generalisasi berlebihan)
4	Rendah	30	22	73.3%	*Jarene* (X), *kanthi* (ADP→SCONJ), *asih* (VERB→ADJ), *pametu* (NOUN→VERB)

Kesalahan pada sistem mencerminkan tantangan linguistik spesifik bahasa Jawa yang belum sepenuhnya teratasi. Pada kategori ADJ (kata sifat) dengan *F1-score* 0,76, model mengalami kesulitan membedakan fungsi ganda kata sifat yang dapat berperan sebagai deskriptor atau nomina tergantung konteks. Contohnya, kata "*apik*" dapat berarti "bagus" (kata sifat) atau merujuk pada "kebaikan" (kata benda), mencerminkan ambiguitas morfologis bahasa Jawa yang kompleks. Pola kesalahan ini menunjukkan bahwa *fine-tuning* BERT, meskipun efektif, belum sepenuhnya menangkap nuansa sosiolinguistik dan konteks panjang-jarak yang diperlukan untuk klasifikasi sempurna pada bahasa dengan variasi gramatikal tinggi.

Hasil pengujian ini membentuk profil kapabilitas model yang cukup jelas, performa sangat baik pada kalimat berpola standar, menurun pada kalimat dengan morfologi derivasional kompleks, dan paling lemah pada kategori gramatikal yang memerlukan pemahaman pragmatik atau konteks lintas-kalimat. Profil ini konsisten dengan arsitektur BERT yang mengoptimalkan representasi kontekstual, kuat untuk dependensi lokal, terbatas untuk inferensi pragmatik. Untuk penggunaan praktis, model ini cocok diterapkan pada teks formal dan kalimat dengan struktur baku, dan memerlukan peningkatan data serta teknik augmentasi untuk menangani variasi informal dan bentuk derivasional yang langka.

IV. KESIMPULAN

Model BERT yang telah melalui proses *fine-tuning* berhasil mengatasi tugas *Part-of-Speech Tagging* untuk bahasa Jawa dengan akurasi 88,87%, secara langsung menjawab tujuan penelitian untuk mengembangkan model POS tagging yang akurat bagi bahasa Jawa. Keberhasilan ini membuktikan bahwa pendekatan *deep learning pre-trained* dapat efektif diterapkan pada bahasa dengan sumber daya terbatas. Kemampuan model dalam mencapai konvergensi stabil tanpa *overfitting* serta performa yang konsisten sepanjang proses pelatihan (dengan perbedaan hanya 0,57% antara validasi dan tes) menunjukkan bahwa *fine-tuning* BERT merupakan strategi yang tepat untuk memanfaatkan pengetahuan linguistik umum pada konteks bahasa khusus. Dibandingkan dengan penelitian sebelumnya, model ini menunjukkan keseimbangan yang lebih baik antara akurasi tinggi dan *robustness*—akurasi 88,87% yang dicapai lebih stabil daripada HMM (92,6%) yang cenderung *overfitting* pada data pelatihan yang mirip, serta lebih baik dari CRF (67%) yang terbatas dalam menangkap konteks jarak jauh dalam bahasa Jawa

Meskipun penelitian mencapai hasil yang memuaskan, terdapat beberapa keterbatasan yang perlu dicatat untuk pengembangan selanjutnya. Pertama, evaluasi model hanya dilakukan pada dataset UD-Javanese-CSUI dengan skala terbatas (1.000 kalimat), sehingga generalisasi terhadap korpus bahasa Jawa lainnya atau domain spesifik belum teruji secara menyeluruh. Kedua, penelitian tidak mengeksplorasi variasi tingkat tutur bahasa Jawa (Ngoko, Krama, dan Krama Inggil) secara seimbang dalam dataset pelatihan, padahal aspek sosiolinguistik ini sangat signifikan dalam konteks penggunaan bahasa Jawa. Ketiga, analisis tentang hubungan antara kompleksitas arsitektur model dan peningkatan akurasi belum dilakukan secara mendalam untuk memahami *trade-off* antara efisiensi dan performa. Terakhir, aspek efisiensi komputasi dalam proses *fine-tunin*g belum dioptimalkan secara komprehensif untuk implementasi pada perangkat dengan komputasi terbatas.

Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk memperluas dan menyeimbangkan dataset dengan memasukkan ketiga variasi tingkat tutur bahasa Jawa untuk meningkatkan *robustness* model terhadap keragaman linguistik, melakukan komparasi sistematis dengan menggunakan model berukuran lebih besar (seperti *base-sized* BERT atau RoBERTa) untuk memahami hubungan antara kompleksitas arsitektur dan peningkatan akurasi, mengimplementasikan teknik optimasi komputasi seperti

quantization, *pruning*, *gradient accumulation*, dan *mixed precision training* untuk efisiensi sumber daya, serta memanfaatkan infrastruktur *cloud computing* untuk skalabilitas, dan menginvestigasi *transfer learning* BERT yang telah di-*fine-tune* untuk tugas NLP terkait pada bahasa Jawa seperti *Named Entity Recognition* (NER) dan *dependency parsing*, dengan potensi implementasi menjadi aplikasi pembelajaran bahasa Jawa interaktif yang berbasis konteks linguistik dan dapat diakses di perangkat *mobile*.

DAFTAR PUSTAKA

[1] S. M. Ah, R. D. Permata, and R. Nugrahani, "Pengaruh Pemanfaatan Aplikasi Digital Berbasis Android terhadap Perkembangan Bahasa Jawa pada Anak Usia Dini," *Indones. Res. J. Educ.*, vol. 5, no. https://irje.org/irje/issue/view/15, pp. 155 – 163, 2025, doi: https://doi.org/10.31004/irje.v5i1.1801.

[2] I. Alfina, A. Yuliawati, D. Tanaya, A. Dinakaramani, and D. Zeman, "A Gold Standard Dataset for Javanese Tokenization, POS Tagging, Morphological Feature Tagging, and Dependency Parsing," *Forum Linguist. Stud.*, vol. 6, no. 5, pp. 131–148, 2024, doi: 10.30564/fls.v6i5.6957.

[3] A. Raup, W. Ridwan, Y. Khoeriyah, S. Supiana, and Q. Y. Zaqiah, "Deep Learning dan Penerapannya dalam Pembelajaran," *JIIP - J. Ilm. Ilmu Pendidik.*, vol. 5, no. 9, pp. 3258–3267, 2022, doi: 10.54371/jiip.v5i9.805.

[4] T. M. Nasir, Y. K. I. Rohima, M. Sabarudin, M. Yasir, S. Supiana, and Q. Y. Zaqiah, "Innovation in the Field of Learning: Deep Learning Approach and Its Application in Learning at Hayat School Bandung City," *Al Ulya J. Pendidik. Islam*, vol. 10, no. 2, pp. 221–239, 2025, doi: 10.32665/alulya.v10i2.4414.

[5] Y. Banua and W. Wiji, "The Implementation of Deep Learning Based Experiential Learning in Developing Metacognitive and Critical Thinking Skills of High School Students: A Systematic Literature Review," *Eurasia Proc. Educ. Soc. Sci.*, vol. 46, pp. 10–19, 2025, doi: 10.55549/epess.977.

[6] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *J. Comput. Cogn. Eng.*, vol. 2, no. 4, pp. 352–356, 2023, doi: 10.47852/bonviewJCCE3202838.

[7] M. Raquib *et al.*, "A Unified BERT-CNN-BiLSTM Framework for Simultaneous Headline Classification and Sentiment Analysis of Bangla News," pp. 1–16, 2025, [Online]. Available: http://arxiv.org/abs/2511.18618

[8] Y. Liang and J. Liu, "Robust Text Classification via Improved CNN, Unbalanced BiLSTM, and Multi-Head Attention," *Informatica*, vol. 49, no. 35, pp. 95–108, 2025, doi: 10.31449/inf.v49i35.11100.

[9] M. Homburg *et al.*, "AI-driven early infectious disease detection in Dutch primary care using BERT and ERNIE," *npj Digit. Med.*, 2025, doi: 10.1038/s41746-025-02278-7.

[10] J. Rawa and J. Sienkiewicz, "Quantifying correlations between information overload and fake news during COVID-19 pandemic: a Reddit study with BERT model approach," pp. 1–22, 2026, [Online]. Available: http://arxiv.org/abs/2601.00496

[11] M. Alfian, U. L. Yuhana, and D. Siahaan, "Indonesian Part-of-Speech Tagger: A Comparative Study," *2023 10th Int. Conf. Adv. Informatics Concept, Theory Appl. ICAICTA 2023*, no. October 2023, pp. 1–6, 2023, doi: 10.1109/ICAICTA59291.2023.10390353.

[12] M. Alfian, U. L. Yuhana, D. Siahaan, H. Munazharoh, and E. Pardede, "Out-of-Vocabulary Handling in Part-of-Speech Tagging: A Semantic Web-Driven Systematic Review," *Int. J. Semant. Web Inf. Syst.*, vol. 21, no. 1, pp. 1–36, 2025, doi: 10.4018/IJSWIS.388421.

[13] A. Sultana and F. Ahmed, "Explicit Grammar Semantic Feature Fusion for Robust Text Classification," 2026, [Online]. Available: http://arxiv.org/abs/2602.20749

[14] A. Zilziana, A. A. Suryani, and I. Asror, "Part of Speech Tagging Menggunakan Bahasa Jawa Dengan Metode Condition Random Fields," *e-Proceeding Eng.*, vol. 7, no. 2, pp. 8103–8111, 2020.

[15] H. Li, H. Mao, and J. Wang, "Part-of-speech tagging with rule-based data preprocessing and transformer," *Electron.*, vol. 11, no. 1, 2022, doi: 10.3390/electronics11010056.

[16] H. Visuwalingam, R. Sakuntharaj, J. Alawatugoda, and R. Ragel, "Deep Learning Model for Tamil Part-of-Speech Tagging," *Comput. J.*, vol. 67, no. 8, pp. 2633–2642, 2024, doi: 10.1093/comjnl/bxae033.

[17] P. Sonawane, K. T. Patil, R. P. Bhavsar, and B. V Pawar, "POS Tagging : A Review of Recent Techniques," 2026.

[18] Ryan Armiditya Pratama, A. A. Suryani, and W. Maharani, "Part of Speech Tagging for Javanese Language with Hidden Markov Model," *J. Comput. Sci. Informatics Eng.*, vol. 4, no. 1, pp. 84–91, 2020, doi: 10.29303/jcosine.v4i1.346.

[19] D. Fimoza, A. Amalia, and T. Henny Febriana Harumy, "Sentiment Analysis for Movie Review in Bahasa Indonesia Using BERT," *2021 Int. Conf. Data Sci. Artif. Intell. Bus. Anal. DATABIA 2021 - Proc.*, pp. 27–34,

2021, doi: 10.1109/DATABIA53375.2021.9650096.

[20] P. You, C. So, S. Choe, and Y. Lee, "Word Embeddings Network and Transformer Based Part of Speech Tagging for Korean," vol. 12, no. 1, pp. 11–24, 2026.

[21] Y. Jumaryadi, R. Meiyanti, R. Fajriah, A. N. Mahsyar, and P. S. Anggraeni, "Implementasi Algoritma Random Forest untuk Analisis Sentimen Ulasan Pengguna Aplikasi Merdeka Mengajar," *Bull. Comput. Sci. Res.*, vol. 5, no. 4, pp. 813–820, 2025, doi: 10.47065/bulletincsr.v5i4.530.

[22] W. Wongso, D. S. Setiawan, and D. Suhartono, "Causal and Masked Language Modeling of Javanese Language using Transformer-based Architectures," *2021 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2021*, 2021, doi: 10.1109/ICACSIS53237.2021.9631331.