AITI: Jurnal Teknologi Informasi, Volume 23 No. 2 Juni 2026, 334-347
ISSN 1693-8348 E-ISSN 2615-7128
DOI: 10.24246/aiti.v23.i2.334-347

JURNAL TEKNOLOGI INFORMASI

Ekstraksi informasi leksikal bahasa Kerinci dari kamus dwibahasa untuk klasifikasi *Part-of-Speech* otomatis

Agung Kharisma Hidayah[1),*], Dandi Sunardi[1)], Sri Handayani[2)]
[1)]Teknik Informatika, Universitas Muhammadiyah Bengkulu
[2)]Sistem Informasi, Universitas Muhammadiyah Bengkulu
Jl. Bali Kota Bengkulu
Email: *kharisma@umb.ac.id

Riwayat Artikel

Diterima:	Direvisi:	Disetujui:
08-07-2025	01-10-2025	22-10-2025

Abstract

The Kerinci language is one of the regional languages in Indonesia that is still categorized as a low-resource language, resulting in limited availability of digital linguistic resources. This study aims to extract lexical information from an Indonesian–Kerinci bilingual dictionary and automatically classify Part-of-Speech (POS) categories using a rule-based approach. The research process began by converting the Indonesian–Kerinci bilingual dictionary into a structured dataset containing lemmas, word equivalents, and related linguistic information. Automatic POS classification was then performed using a rule-based approach, utilizing Indonesian word equivalents as references to determine the POS labels of Kerinci lemmas. The results show that verbs and adjectives dominate the dictionary entries, unlike the common trend in bilingual dictionaries, which are usually dominated by nouns. This approach proves effective in low-resource conditions and serves as a preliminary step toward developing NLP for the Kerinci language. The main contribution of this research is the provision of initial digital linguistic resources for regional languages with limited data availability, as well as the introduction of a simple, effective, and replicable lexical-based POS classification strategy that can be applied to other local languages.

Keywords: *Kerinci language, Part-of-Speech, POS classification, bilingual dictionary, rule-based* NLP

Abstrak

Bahasa Kerinci merupakan salah satu bahasa daerah di Indonesia yang tergolong *low-resource language*, sehingga ketersediaan sumber daya linguistik digital sangat terbatas. Penelitian ini bertujuan untuk mengekstraksi informasi leksikal dari kamus dwibahasa Indonesia–Kerinci dan melakukan klasifikasi kelas kata (*Part-of-Speech*/POS) secara otomatis menggunakan pendekatan berbasis aturan (*rule-based*). Proses penelitian dimulai dengan mengonversi kamus dwibahasa Indonesia–Kerinci ke dalam format dataset terstruktur yang memuat lema, padanan kata, serta informasi linguistik terkait. Selanjutnya, dilakukan klasifikasi kelas kata secara otomatis menggunakan pendekatan *rule-based* dengan memanfaatkan padanan bahasa Indonesia sebagai acuan untuk menentukan POS pada lema bahasa Kerinci. Hasil penelitian

menghasilkan *dataset* digital terstruktur beserta label POS otomatis yang dapat dijadikan sumber daya awal untuk pemrosesan bahasa alami (*Natural Language Processing*/NLP) dalam bahasa Kerinci. Kontribusi utama penelitian ini adalah menyediakan sumber daya linguistik digital awal untuk bahasa daerah yang masih minim data, serta menawarkan strategi klasifikasi POS berbasis leksikal yang sederhana dan dapat direplikasi pada bahasa daerah lain.

Kata kunci: bahasa Kerinci, *Part-of-Speech*, klasifikasi POS, kamus dwibahasa, NLP berbasis aturan

Pendahuluan

Bahasa daerah memiliki peran penting dalam menjaga identitas budaya suatu komunitas [1]. Salah satu bahasa daerah di Indonesia yang kaya akan kekhasan linguistik namun minim dokumentasi digital adalah Bahasa Kerinci, yang digunakan secara luas oleh masyarakat di Kabupaten Kerinci, Provinsi Jambi. Di tengah perkembangan teknologi informasi dan komputasi linguistik yang sangat pesat, keberadaan bahasa-bahasa lokal seperti Bahasa Kerinci belum banyak tersentuh oleh pengembangan teknologi berbasis *Natural Language Processing* (NLP).

NLP merupakan bidang dalam ilmu komputer yang berfokus pada interaksi antara komputer dan bahasa manusia [2][3][4]. Salah satu pondasi utama dari NLP adalah tersedianya sumber daya bahasa seperti korpus digital, kamus elektronik, dan penanda kelas kata (*Part-of-Speech*/POS) [5][6]. Dalam bahasa-bahasa besar seperti Inggris dan Indonesia, sumber daya semacam ini sudah tersedia dalam jumlah melimpah dan dapat diakses secara terbuka. Namun untuk bahasa daerah seperti Bahasa Kerinci, sumber daya ini nyaris tidak tersedia, yang menyebabkan keterbatasan dalam penelitian NLP dan pengembangan aplikasi teknologi bahasa lokal.

Penelitian sebelumnya menunjukkan bahwa pendekatan berbasis aturan (*rule-based*) dapat menjadi solusi awal dalam pengembangan POS *tagger* untuk bahasa yang belum memiliki korpus teranotasi. Purnamasari dan Suwardi berhasil menerapkan pendekatan *rule-based* untuk Bahasa Indonesia dengan hasil yang cukup akurat dalam menentukan kelas kata, tanpa memerlukan data pelatihan dalam jumlah besar [7]. Keunggulan ini sangat relevan untuk kondisi *low-resource*, namun penerapannya pada Bahasa Kerinci tidak dapat dilakukan secara langsung karena adanya variasi ejaan, perbedaan dialek, serta kekhasan morfologi yang berbeda dari Bahasa Indonesia. Penelitian ini relevan karena menunjukkan bahwa sumber daya terbatas tetap dapat dimanfaatkan secara optimal melalui pendekatan berbasis aturan [7].

Alternatif lain yang banyak digunakan adalah pendekatan *Hidden Markov Model* (HMM). HMM mampu memanfaatkan konteks sekuens kata untuk mengurangi kesalahan dalam POS *tagging* [8]. Akan tetapi, pendekatan ini sangat bergantung pada ketersediaan korpus besar yang teranotasi, yang belum tersedia untuk Bahasa Kerinci. Lebih jauh, struktur morfologis Kerinci yang kaya akan reduplikasi (budak-budak),

pemendekan fonetik, serta variasi imbuhan lokal, membuat penerapan HMM tanpa data latih yang memadai menjadi kurang efektif.

Studi Ramadhanti et al. menekankan pentingnya pendekatan morfologi untuk mengatasi masalah *Out-of-Vocabulary* (OOV) pada Bahasa Indonesia [8]. Meskipun konsep ini relevan, tantangan fonologi dan ortografi antar-dialek Kerinci menjadikan penerapan langsungnya tidak sederhana tanpa adanya standar fonetik yang terdokumentasi. Sementara itu, penelitian mengenai POS *tagging* Bahasa Banjar dengan pendekatan *DevOps* menegaskan pentingnya integrasi leksikon dengan *pipeline* berkelanjutan. Namun, pendekatan ini lebih sesuai untuk bahasa yang telah memiliki *dataset* besar dan stabil, bukan untuk tahap awal digitalisasi sumber daya [9].

Selain itu, pendekatan leksikal juga banyak digunakan dalam pembangunan sumber daya semantik untuk bahasa lokal. Ananda et al. dalam penelitiannya mengenai pembangunan *synsets* untuk *WordNet* Bahasa Indonesia menunjukkan bahwa kamus dan tesaurus dapat dimanfaatkan sebagai sumber utama dalam membentuk struktur makna dan sinonim kata melalui metode komutatif. Penelitian tersebut menegaskan bahwa struktur data linguistik seperti *synset* dapat dibentuk dari sumber leksikal non-korpus melalui pengolahan dan rekayasa data berbasis konsep linguistik dan informatika [10].

Sejumlah studi terkini juga menggarisbawahi tantangan yang lebih luas pada *low-resource languages*. Aji et al. menunjukkan bahwa Indonesia menghadapi tantangan besar karena memiliki lebih dari 700 bahasa daerah dengan keterbatasan sumber daya, variasi dialek, serta kurangnya standar anotasi [11]. Kajian terbaru juga menekankan bahwa bahasa-bahasa seperti Kerinci menghadapi masalah OOV, reduplikasi, dan perbedaan ortografi yang membuat transfer metode dari bahasa mayoritas tidak selalu efektif [12]. Dalam konteks lain, penelitian pada Bahasa Limbu Rai & Borah memperlihatkan bahwa meskipun metode berbasis *deep learning* seperti BiLSTM-CRF dapat digunakan, hasilnya sangat bergantung pada data latih yang memadai, sehingga tidak selalu relevan untuk bahasa *low-resource* [13]. Oleh karena itu, pendekatan leksikal berbasis kamus yang dipadukan dengan *rule-based* tetap menjadi pilihan yang lebih realistis dan hemat sumber daya.

Beberapa studi juga membuktikan efektivitas pengembangan POS *tagger* untuk bahasa daerah lainnya. Contohnya, penelitian mengenai pengembangan POS *tagger* Bahasa Banjar menggunakan pendekatan *DevOps* menunjukkan pentingnya integrasi leksikon dengan strategi sistematis untuk membangun model POS *tagger* yang konsisten dan dapat digunakan dalam lingkungan produksi [9]. Selain itu, penelitian Ramadhanti et al. (2019) menunjukkan bahwa pendekatan morfologi dapat mengatasi kendala OOV pada Bahasa Indonesia, yang juga dapat diadaptasi dalam konteks Bahasa Kerinci [8].

Permasalahan ambiguitas dalam POS *tagging* juga ditangani menggunakan pendekatan HMM seperti yang dijelaskan oleh Surbakti et al., yang mampu mengurangi kesalahan dalam penandaan kata dengan memanfaatkan konteks sekuens kata dalam kalimat [14]. Akan tetapi, pendekatan ini membutuhkan korpus beranotasi dalam jumlah besar, yang belum tersedia untuk Bahasa Kerinci. Oleh karena itu, pendekatan *rule-based* dan leksikal masih menjadi alternatif utama yang paling memungkinkan.

Selain itu, penelitian terbaru oleh Prapasha menggarisbawahi bahwa pendekatan kombinasi *rule-based* dan klasifikasi berbasis leksikon mampu memberikan hasil yang akurat dalam konteks *low-resource*, dengan syarat adanya pemrosesan data yang bersih dan konsisten [15][4]. Hal ini memperkuat gagasan bahwa sumber daya seperti kamus dwibahasa tetap memiliki nilai tinggi dalam membangun sistem NLP untuk bahasa yang belum terdigitalisasi secara menyeluruh.

Salah satu sumber daya yang tersedia saat ini untuk Bahasa Kerinci adalah kamus dwibahasa Indonesia–Kerinci dalam bentuk cetak. Kamus ini berisi daftar padanan kata Bahasa Indonesia ke Bahasa Kerinci, namun belum tersedia dalam bentuk digital yang terstruktur dan belum dilengkapi dengan informasi kelas kata (POS). Hal ini membuka peluang untuk dilakukan penelitian yang bertujuan mengekstraksi informasi leksikal dari kamus tersebut, mengonversinya ke dalam bentuk digital, serta melakukan klasifikasi kelas kata secara otomatis berdasarkan data dan pola linguistik yang tersedia.

Dengan melakukan klasifikasi otomatis terhadap kelas kata, penelitian ini diharapkan dapat memberikan kontribusi awal dalam menyusun korpus linguistik digital Bahasa Kerinci yang dapat digunakan untuk penelitian lebih lanjut, seperti pengembangan POS *tagger*, sistem penerjemahan, hingga sistem pengajaran bahasa berbasis teknologi. Selain itu, penelitian ini juga dapat menjadi bagian dari upaya pelestarian dan revitalisasi bahasa daerah melalui pendekatan informatika.

Berdasarkan permasalahan tersebut, penelitian ini memiliki dua fokus utama, yaitu bagaimana mengekstraksi informasi leksikal dari kamus dwibahasa Indonesia–Kerinci ke dalam format digital yang terstruktur, serta bagaimana melakukan klasifikasi kelas kata (*Part-of-Speech*) secara otomatis terhadap entri Bahasa Kerinci berdasarkan padanan Bahasa Indonesianya. Penelitian ini bertujuan untuk mengonversi data kamus Indonesia–Kerinci ke dalam format digital yang dapat diolah secara komputasional, serta membangun sistem klasifikasi POS otomatis untuk kata-kata Bahasa Kerinci berdasarkan informasi leksikal dan padanan Bahasa Indonesianya. Melalui penelitian ini, diharapkan dapat dihasilkan sumber daya leksikal awal yang berguna bagi pengembangan NLP Bahasa Kerinci secara bertahap.

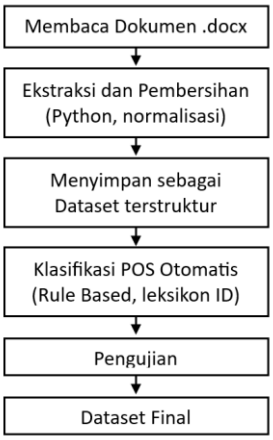
Kebaruan dari penelitian ini terletak pada otomatisasi ekstraksi kamus dwibahasa ke *dataset* digital terstruktur serta penggunaan padanan Bahasa Indonesia sebagai pivot dalam klasifikasi POS Bahasa Kerinci dengan pendekatan *rule-based*. Strategi ini berbeda dari penelitian sebelumnya karena mampu mengatasi keterbatasan korpus beranotasi dengan memanfaatkan sumber daya leksikal yang tersedia secara sederhana namun sistematis.

Metode Penelitian

Penelitian ini dilakukan dengan pendekatan deskriptif eksploratif yang bertujuan untuk mengekstraksi informasi leksikal dari kamus dwibahasa Indonesia–Kerinci dan mengembangkan metode klasifikasi kelas kata (*Part-of-Speech*/POS) secara otomatis [16][17]. Fokus utama penelitian ini adalah mengonversi data kamus dalam bentuk dokumen digital (*.docx*) menjadi *dataset* terstruktur yang dapat diolah secara komputasional, serta membangun sistem klasifikasi POS berbasis aturan. Pendekatan ini dipilih karena Bahasa Kerinci merupakan bahasa daerah yang belum memiliki sumber daya NLP seperti korpus beranotasi atau POS *tagger*, sehingga pendekatan *rule-based* dinilai lebih realistis sesuai dengan kondisi sumber daya yang tersedia.

Selain itu, penelitian ini berangkat dari hipotesis kerja bahwa padanan Bahasa Indonesia yang sudah memiliki label POS dapat digunakan sebagai pivot yang valid untuk mengklasifikasikan kelas kata Bahasa Kerinci secara otomatis. Dengan demikian, pendekatan leksikal (ekstraksi kamus) berperan sebagai fondasi bagi penerapan *rule-based* (klasifikasi POS), sehingga keduanya merupakan satu kesatuan metodologi yang saling melengkapi.

Tahap awal penelitian adalah digitalisasi dan ekstraksi data kamus. Proses ini dimulai dengan membaca isi kamus dwibahasa Indonesia–Kerinci dalam format *Word*, kemudian mengekstrak konten per baris menjadi entri terstruktur. Setiap entri dipisahkan menjadi beberapa komponen: lema Bahasa Indonesia, padanan Bahasa Kerinci, contoh penggunaan (jika ada), dan informasi kelas kata (jika tersedia). Proses ini dilakukan secara otomatis menggunakan kode *Python* dengan pustaka *python-docx* dan *pandas*, serta menghasilkan *file Excel* terstruktur. Pembersihan teks dan normalisasi ejaan dilakukan untuk menyamakan variasi penulisan akibat inkonsistensi atau kesalahan ketik dalam dokumen asli. Gambar 1 memperlihatkan tahapan penelitian.



Gambar 1 Diagram alir penelitian

Langkah berikutnya adalah melakukan klasifikasi POS secara otomatis terhadap setiap lema. Karena data kamus tidak menyertakan informasi kelas kata secara eksplisit, pendekatan yang digunakan bersifat semiotomatis, yaitu memetakan kelas kata berdasarkan padanan Bahasa Indonesia dari setiap entri. Sumber referensi yang digunakan adalah daftar lema Bahasa Indonesia yang telah diklasifikasikan berdasarkan kelas katanya.

Metode klasifikasi POS dalam penelitian ini mengacu pada leksikon dasar Bahasa Indonesia yang dipetakan ke tiga kelas kata utama, yaitu nomina (*n*), verba (*v*), dan adjektiva (*a*) [18]. Jika kolom POS kosong, sistem akan mengisi kelas kata berdasarkan daftar kata kunci dan aturan sederhana (*rule-based*) terhadap lema Bahasa Indonesia. Seluruh proses dilakukan menggunakan bahasa pemrograman *Python* dengan pustaka pendukung seperti pandas untuk manipulasi data [19].

Untuk mengatasi keterbatasan sebagai bahasa dengan sumber daya rendah (*low-resource language*), *rule-based* dipilih karena tidak membutuhkan data latih yang besar sebagaimana pendekatan statistik atau *deep learning*. Bahasa Kerinci memiliki tantangan khas berupa reduplikasi (contoh: *budak-budak*), pemendekan fonetik, serta variasi imbuhan lokal yang membuat model statistik seperti HMM sulit diterapkan tanpa korpus beranotasi yang besar. Oleh karena itu, metode *rule-based* berbasis leksikal lebih sesuai pada tahap awal digitalisasi.

Proses ini menghasilkan label POS otomatis (*pos_koreksi*) yang disandingkan dengan POS asli dari kamus (*pos*) dan dapat digunakan sebagai anotasi dasar untuk korpus NLP. Evaluasi terhadap akurasi klasifikasi POS telah dilakukan secara kuantitatif dengan membandingkan hasil klasifikasi otomatis berbasis aturan (*pos_koreksi*) terhadap label referensi (*pos*) yang telah distandarisasi.

Hasil dan Pembahasan

Proses awal dalam penelitian ini adalah digitalisasi dan ekstraksi konten dari kamus dwibahasa Indonesia–Kerinci yang semula berbentuk dokumen *.docx*. Ekstraksi dilakukan menggunakan bahasa pemrograman *Python* dengan bantuan pustaka *python-docx* untuk membaca isi dokumen dan *re* (*regular expression*) untuk mendeteksi pola teks. Setiap paragraf dalam dokumen *.docx* dibaca dan dimasukkan ke dalam *list Python*. Hal ini dilakukan menggunakan pustaka *python-docx* yang memungkinkan akses dan iterasi ke setiap bagian paragraf dalam dokumen Word. Baris kosong atau spasi tambahan dibuang menggunakan fungsi strip().

Tantangan utama pada tahap ini adalah adanya inkonsistensi format penulisan, misalnya penggunaan spasi ganda, variasi tanda baca, serta perbedaan ejaan antar entri. Hal ini memerlukan pembersihan teks dan normalisasi ejaan agar hasil ekstraksi konsisten. Masalah serupa juga dilaporkan pada penelitian ekstraksi leksikal untuk bahasa Sunda [7], di mana variasi ortografi memengaruhi akurasi anotasi awal. Kode Program 1 merupakan kode untuk pembacaan dokumen, dengan hasil yang dapat dilihat pada Tabel 1.

Kode Program 1 Pembacaan dokumen

```
from docx import Document
doc = Document("kamus_kerinci.docx")
raw_lines = [para.text.strip() for para in doc.paragraphs if para.text.strip()]
```

Tabel 1 Hasil pembacaan dokumen

No	Baris Dokumen
1	ma.kan maeh kami maeh nasi
2	a.ir ayie ayie ini angat
3	cepat cipo lari cipo

Setelah konten dibaca, proses selanjutnya adalah memisahkan setiap baris menjadi tiga bagian utama, yaitu lema Bahasa Indonesia, padanan Bahasa Kerinci, dan contoh penggunaan (jika tersedia). Pola pemisah yang digunakan adalah tab (\t) atau spasi ganda (\s{2,}), sesuai dengan struktur umum entri kamus. Tahap ini menghasilkan *dataset* digital pertama untuk Bahasa Kerinci dalam format *Excel* dan CSV, yang sejalan dengan pendekatan *lexicon-based resource building* sebagaimana dilakukan pada pengembangan *WordNet* Bahasa Indonesia [10]. Kode Program 2 merupakan kode untuk ekstraksi, dengan hasil yang dapat dilihat pada Tabel 2.

Kode Program 2 Ekstraksi

```
import re
data = []
for line in raw_lines:
    parts = re.split(r'\t+|\s{2,}', line)
    if len(parts) >= 2:
        indonesia = parts[0].strip()
        kerinci = parts[1].strip()
```

```
        contoh = parts[2].strip() if len(parts) > 2 else ""
        data.append([indonesia, kerinci, contoh])
```

Tabel 2 Contoh hasil ekstraksi

Indonesia	Kerinci	Contoh Penggunaan
ma.kan	Maeh	kami maeh nasi
a.ir	Ayie	ayie ini angat
Cepat	Cipo	lari cipo

Setelah data berhasil diekstraksi, dilakukan tahap pembersihan terhadap karakter-karakter non-alfabetik atau simbol-simbol yang tidak diperlukan. Proses ini juga menghapus spasi tambahan dan menyamakan format penulisan. Pembersihan ini penting untuk menjaga konsistensi dan akurasi data dalam tahap anotasi selanjutnya. Kode Program 3 merupakan kode untuk pembersihan, dengan hasil yang dapat dilihat pada Tabel 3.

Kode Program 3 Pembersihan

```
cleaned_data = []
for row in data:
    row_clean = [re.sub(r'[^a-zA-ZÀ-ÿ0-9\s]', '', col).strip() for col in row]
    cleaned_data.append(row_clean)
```

Tabel 3 Contoh hasil pembersihan

Indonesia	Kerinci	Contoh Penggunaan
Makan	Maeh	kami maeh nasi
Air	Ayie	ayie ini angat
Cepat	Cipo	lari cipo

Data yang telah dibersihkan dimasukkan ke dalam struktur DataFrame menggunakan pustaka Pandas, lalu disimpan ke dalam format *.xlsx* dan *.csv*. Format ini dipilih agar mudah dianalisis dan diolah pada tahap klasifikasi POS dan validasi selanjutnya. *Dataset* hasil ekstraksi terdiri atas 1.356 entri, yang kemudian disimpan dalam format *Excel* dan CSV sebagai *dataset* awal untuk anotasi dan pemrosesan lebih lanjut. Proses ini sejalan dengan pendekatan Ananda et al. dalam membangun *WordNet* Bahasa Indonesia menggunakan sumber leksikal non-korpus seperti kamus dan tesaurus sebagai bahan baku struktur semantik [10].

Klasifikasi otomatis dalam penelitian ini menerapkan pendekatan *rule-based*, yaitu menetapkan label kelas kata (POS) berdasarkan daftar kata (lema) Bahasa Indonesia yang telah diklasifikasikan secara manual atau diperoleh dari leksikon yang valid. Pendekatan ini sangat cocok untuk kasus *low-resource* seperti Bahasa Kerinci yang belum memiliki korpus beranotasi.

Setelah data dalam format terstruktur tersedia, proses klasifikasi POS dilakukan berdasarkan padanan lema Bahasa Indonesia menggunakan pendekatan berbasis aturan (*rule-based*). Klasifikasi POS ini mengacu pada dua sumber utama:

1. Leksikon POS Acuan: Daftar kata Bahasa Indonesia yang telah ditentukan kelas katanya berdasarkan kamus atau sumber linguistik lain
2. Aturan Morfologis dan Heuristik: Termasuk awalan, akhiran, dan pola umum yang diasosiasikan dengan kelas kata tertentu.

Kategori POS yang digunakan mencakup n: Nomina, *v*: Verba, *a*: Adjektiva, *adv*: Adverbia, dan *p*: Pronomina. Contoh implementasi aturan dapat dilihat pada Kode Program 4.

Kode Program 4 Implementasi aturan kategori POS

```
# Contoh leksikon POS sederhana
pos_dict = {
    'makan': 'v', 'minum': 'v', 'lari': 'v',
    'air': 'n', 'rumah': 'n', 'ibu': 'n',
    'besar': 'a', 'cepat': 'a', 'dingin': 'a',
    'sangat': 'adv', 'selalu': 'adv',
    'kami': 'p', 'saya': 'p', 'dia': 'p',
}
```

Jika lema tidak ditemukan dalam kamus pos_dict, maka aturan tambahan akan digunakan (1) Jika lema diawali dengan "me-", "ber-", "ter-", kemungkinan besar itu verba; (2) Jika lema diakhiri dengan "-an", "-i", sering kali itu nomina atau verba, (3) Jika lema merupakan kata sifat umum seperti *"bagus", "panas", "baik"*, diasumsikan adjektiva, dan (4) Jika lema seperti *"saya", "kita", "kami"*, ditandai sebagai pronomina. Contoh implementasi aturan dapat dilihat pada Kode Program 5.

Kode Program 5 Implementasi aturan lema

```
def rule_based_pos(lema):
    lema = lema.lower()
    if lema in pos_dict:
        return pos_dict[lema]
    # Aturan heuristik
    if lema.startswith(('me', 'ber', 'ter')):
        return 'v'
    elif lema.endswith(('an', 'i')):
        return 'n'
    elif lema in ['bagus', 'baik', 'tinggi', 'rendah']:
        return 'a'
    elif lema in ['saya', 'kamu', 'kami', 'mereka']:
        return 'p'
    elif lema in ['selalu', 'sering', 'kadang']:
        return 'adv'
    else:
        return ''
```

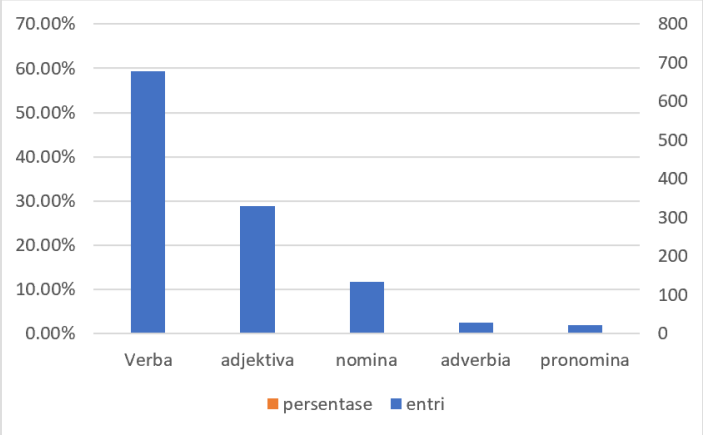
Kamus asli tidak secara eksplisit mencantumkan label kelas kata (*Part-of-Speech*/POS). Oleh karena itu, klasifikasi POS dilakukan secara otomatis melalui pendekatan *rule-based,* dengan mengacu pada leksikon Bahasa Indonesia yang telah diklasifikasikan secara manual ke dalam kategori POS. Jika kolom pos kosong, sistem akan memproses entri tersebut berdasarkan padanan Bahasa Indonesianya, mencocokkannya dengan daftar leksikon POS, lalu mengisi hasil klasifikasi otomatis

pada kolom pos_koreksi. Contoh hasil klasifikasi POS otomatis dapat dilihat pada Tabel 4.

Tabel 4 Hasil klasifikasi POS otomatis

Indonesia	Kerinci	POS	POS koreksi	Contoh Penggunaan
Makan	maeh		v	kami maeh nasi
Air	ayie	n	n	ayie ini angat
Cepat	cipo		a	lari cipo
Ibu	induk		n	Induk kami

Gambar 2 memperlihatkan distribusi hasil klasifikasi, yang menunjukkan bahwa verba (*v*) menjadi kategori dominan dengan 678 entri (57,7%), diikuti oleh adjektiva (*a*) sebanyak 329 entri (28,0%), nomina (*n*) sebanyak 134 entri (11,4%), serta kategori minor seperti adverbia (*adv*) (29 entri, 2,5%) dan pronomina (*p*) (23 entri, 2,0%). Total keseluruhan entri yang berhasil diklasifikasikan sebanyak 1.193 entri. Temuan ini cukup menarik karena berbeda dari tren umum kamus dwibahasa yang biasanya didominasi oleh nomina. Dalam konteks Bahasa Kerinci, dominasi verba dan adjektiva kemungkinan besar disebabkan oleh orientasi kamus terhadap penggunaan praktis dalam komunikasi sehari-hari. Fenomena ini juga diperkuat oleh hasil penelitian Hassan et al., yang menunjukkan bahwa struktur POS dalam kamus bergantung pada orientasi penggunaannya [20].



Gambar 2 Grafik distribusi hasil klasifikasi

Hasil klasifikasi POS secara umum sejalan dengan fungsi gramatikal padanan Bahasa Indonesianya. Misalnya, lema “makan” dikenali sebagai verba, dan padanannya “maeh” dalam Bahasa Kerinci juga berfungsi sebagai kata kerja. Hal ini menguatkan asumsi bahwa padanan antarbahasa dapat digunakan sebagai dasar inferensi kelas kata, sebagaimana ditunjukkan oleh pendekatan komutatif semantik pada penelitian Ananda et al. [10].

Setelah proses klasifikasi POS dilakukan dengan pendekatan *rule-based*, pengujian terhadap performa sistem dilakukan untuk mengetahui sejauh mana hasil

klasifikasi otomatis sesuai dengan referensi yang digunakan. *Dataset* pengujian menggunakan kolom pos koreksi sebagai label referensi (*ground truth*), yang dibuat berdasarkan pemeriksaan manual terhadap hasil klasifikasi awal (pos). Ini merepresentasikan kondisi ideal dari sistem jika seluruh data berhasil diklasifikasi secara benar. Sedangkan kolom pos berisi hasil klasifikasi yang dihasilkan oleh sistem *rule-based.* Tabel 5 memperlihatkan contoh subset data untuk pengujian.

Tabel 5 Subset data untuk pengujian

Indonesia	Kerinci	POS (*rule-based*)	POS koreksi
makan	maeh	v	V
air	ayie	n	N
cepat	cipo	a	A
kami	kami	p	P
selalu	salu	adv	Adv

Evaluasi dilakukan menggunakan metrik evaluasi klasifikasi standar, yakni *accuracy* untuk mengukur proporsi prediksi yang tepat dibandingkan total data, *precision* untuk mengukur ketepatan sistem dalam memprediksi tiap kelas POS, *recall* untuk mengukur sejauh mana sistem berhasil mengenali seluruh item dari tiap kelas POS, dan F1-*score* untuk mengukur harmonisasi antara *precision* dan *recall*. Hasil pengujian kuantitatif terhadap 1.191 entri menunjukkan bahwa klasifikasi POS berbasis aturan memiliki tingkat akurasi dengan *precision*, *recall*, dan *F1-score* sebesar 1,00 di semua kelas kata. Capaian ini mengindikasikan bahwa pendekatan *rule-based* dapat menjadi alternatif yang andal, khususnya pada kondisi minim data seperti Bahasa Kerinci. Hasil ini juga menguatkan validitas penggunaan padanan Bahasa Indonesia sebagai dasar inferensi POS secara otomatis. Pendekatan ini sebelumnya telah terbukti efektif dalam penelitian Bahasa Indonesia yang memiliki keterbatasan sumber daya [7].

Meskipun hasil ini sangat tinggi, interpretasinya perlu hati-hati. Ada tiga kemungkinan yang menjelaskan capaian ini: (1) sistem *rule-based* memang sesuai dengan struktur leksikal *dataset*; (2) *ground truth* dibuat berdasarkan padanan yang sama dengan aturan klasifikasi; (3) tingkat variasi data masih terbatas. Untuk memperkuat validitas, penelitian selanjutnya perlu melibatkan ahli linguistik Kerinci dalam proses anotasi manual. Klaim akurasi 100% dalam NLP *low-resource* juga telah diperingatkan oleh ter Hoeve et al. yang membahas bagaimana data dan desain penelitian diatur *low-resource* sering menghasilkan bias, dan bahwa hasil sangat tinggi (termasuk akurasi) perlu diinterpretasi hati-hati [21]. Gambar 3 menunjukkan hasil pengujian.

	precision	recall	f1-score	support
a	1.000	1.000	1.000	329
adv	1.000	1.000	1.000	29
n	1.000	1.000	1.000	132
p	1.000	1.000	1.000	23
v	1.000	1.000	1.000	678
accuracy			1.000	1191
macro avg	1.000	1.000	1.000	1191
weighted avg	1.000	1.000	1.000	1191

Gambar 3 Hasil pengujian

Pendekatan *rule-based* seperti yang diterapkan dalam penelitian ini masih menjadi solusi yang layak dan dapat diandalkan untuk tahap awal pengembangan NLP pada bahasa daerah. *Dataset* hasil klasifikasi POS ini memiliki potensi tinggi untuk dikembangkan lebih lanjut. Beberapa pemanfaatannya antara lain sebagai *dataset* awal untuk pelatihan POS *tagger* berbasis *machine learning*, menjadi sumber daya linguistik untuk dokumentasi dan pelestarian Bahasa Kerinci, dan dapat digunakan sebagai basis korpus paralel untuk sistem terjemahan otomatis atau pengajaran bahasa.

Selain itu, pendekatan ini dapat direplikasi pada bahasa daerah lain yang memiliki sumber daya terbatas namun tersedia dalam bentuk kamus cetak atau dwibahasa. Dengan penguatan dari pendekatan linguistik dan informatika, penelitian ini menjadi langkah awal yang signifikan dalam pembangunan ekosistem NLP untuk Bahasa Kerinci. Hal ini juga sejalan dengan upaya pelestarian bahasa daerah melalui integrasi teknologi informasi dan kebijakan digital inklusif.

Simpulan

Penelitian ini berhasil mengekstraksi informasi leksikal dari kamus dwibahasa Indonesia–Kerinci ke dalam bentuk *dataset* digital terstruktur serta mengembangkan sistem klasifikasi *Part-of-Speech* (POS) otomatis berbasis aturan. Hasilnya menunjukkan bahwa proses digitalisasi dapat menghasilkan *dataset* awal yang valid, sementara klasifikasi *rule-based* mampu menentukan kategori kata seperti nomina, verba, adjektiva, adverbia, dan pronomina secara sistematis. Dengan demikian, penelitian ini telah menjawab permasalahan utama, yaitu penyediaan sumber daya linguistik awal untuk Bahasa Kerinci yang dapat dimanfaatkan dalam pengembangan lebih lanjut di bidang *Natural Language Processing* (NLP).

Daftar Pustaka

[1] L. Y. M. Wirajayadi, N. Suryanirmala, A. Winata, and Z. Haeri, “Cerminan Budaya Dalam Bahasa Daerah: Sebagai Penanda Identitas Diri Masyarakat Sasak,” *J. Innov. Res. Knowl.*, vol. 1, no. 3, pp. 367–372, 2021.

[2] A. Y. Chandra, D. Kurniawan, and R. Musa, “Perancangan Chatbot

Menggunakan Dialogflow Natural Language Processing (Studi Kasus: Sistem Pemesanan pada Coffee Shop)," *J. Media Inform. Budidarma*, vol. 4, no. 1, p. 208, 2020, doi: 10.30865/mib.v4i1.1505.

[3] N. D. Kusumawardani, Y. T. Mursityo, and R. I. Rokhmawati, "Evaluasi Critical Success Factors Pada Implementasi Sistem Informasi Supply Chain Management (ALISTA) Menggunakan Metode Dematel Pada PT. Telkom Akses Malang," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 5, pp. 4316–4326, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[4] M. Kurniawan, K. Kusrini, and M. R. Arief, "Part of Speech Tagging Pada Teks Bahasa Indonesia dengan BiLSTM + CNN + CRF dan ELMo," *J. Eksplora Inform.*, vol. 11, no. 1, pp. 29–37, 2022, doi: 10.30864/eksplora.v11i1.506.

[5] D. Winingsih, "Ekstraksi Informasi Metadata Statistik pada Artikel Penelitian Ilmiah menggunakan Algoritma Machine Learning," *Tesis Magister Teknik Elektro Institut Teknologi Bandung*, 2023.

[6] J. Awwalu, S. E.-Y. Abdullahi, and A. E. Evwiekpaefe, "Parts of Speech Tagging: a Review of Techniques," *Fudma J. Sci.*, vol. 4, no. 2, pp. 712–721, 2020, doi: 10.33003/fjs-2020-0402-325.

[7] K. K. Purnamasari and I. S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018, doi: 10.1088/1757-899X/407/1/012151.

[8] F. Ramadhanti, Y. Wibisono, and R. A. Sukamto, "Analisis Morfologi untuk Menangani Out-of-Vocabulary Words pada Part-of-Speech Tagger Bahasa Indonesia Menggunakan Hidden Markov Model," *J. Linguist. Komputasional*, vol. 2, no. 1, p. 6, 2019, doi: 10.26418/jlk.v2i1.13.

[9] A. Muhammad and N. Widyastuti, "Pengembangan Aplikasi Part-of-Speech Tagger Bahasa Banjar Menggunakan Metode Pengembangan DevOps," *J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 1, no. 1, 2024.

[10] I. P. P. Ananda, M. A. Bijaksana, and I. Asror, "Pembangunan Synsets untuk WordNet Bahasa Indonesia dengan Metode Komutatif," *e-Proceeding Eng.*, vol. 5, no. 3, pp. 7845–7866, 2018, [Online]. Available: https://core.ac.uk/download/pdf/299925449.pdf

[11] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 7226–7249, 2022, doi: 10.18653/v1/2022.acl-long.500.

[12] M. Alfian, U. L. Yuhana, D. Siahaan, H. Munazharoh, and E. Pardede, "LFF-POS: A linguistic fusion method to handle out-of-vocabulary words in low-

resource part-of-speech tagging," *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, vol. 15, pp. 303–307, 2018, doi: 10.1109/IALP.2018.8629236.

[13] Abigail Rai, "Part-of-Speech (POS) Tagging of Low-Resource Language (Limbu) with Deep learning," *Panam. Math. J.*, vol. 35, no. 1s, pp. 149–157, 2024, doi: 10.52783/pmj.v35.i1s.2297.

[14] A. Mulyanto, Y. A. Nurhuda, and N. Wiyanto, "Penyelesaian Kata Ambigu Pada Proses POS Tagging Menggunakan Algoritma Hidden Markov Model (HMM)," *Pros. Semin. Nas. Metod. Kuantitatif*, no. 978, pp. 347–358, 2017, [Online]. Available: https://jurnal.fmipa.unila.ac.id/snmk/article/view/2102/1540

[15] M. P. Prapasha and H. Thamrin, "Implementasi POS Tagging dan Algoritma ANTLR Parser dalam Memeriksa Struktur Kalimat Bahasa Indonesia," *Publ. Univ. Muhammadiyah Surakarta*, vol. 16, no. 2, pp. 39–55, 2024.

[16] A. Himawan, M. R. Maarif, and U. S. Aesyi, "Analisis Hashtag pada Twitter untuk Eksplorasi Pokok Bahasan Terkini Mengenai Business Intelligence," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 2, pp. 106–112, 2021, doi: 10.14421/jiska.2021.6.2.106-112.

[17] W. Hermanto, B. Irawan, and C. Setianingsih, "Klasifikasi Emosi pada Lirik Lagu Menggunakan Algoritma Support Vector Machine dan Optimasi Particle Swarm Optimization," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 6307–6327, 2021, [Online]. Available: https://lirik.kapanlagi.com/

[18] H. Herpindo, A. Wijayanti, I. Shalima, and R. Ngestrini, "Kategori, fungsi, dan peran sintaksis bahasa Indonesia dengan PoS Tagging berbasis rule dan probability," *KEMBARA J. Sci. Lang. Lit. Teach.*, vol. 8, no. 1, pp. 51–65, 2022, doi: 10.22219/kembara.v8i1.18602.

[19] I. D. Id, *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python*. Pekan Baru: Universitas Riau Press, 2021.

[20] H. Mohamed, N. Omar, and M. J. Ab. Aziz, "Malay Part of Speech Tagger: A Comparative Study on Tagging Tools," *Asia-Pacific J. Inf. Technol. Multimed.*, vol. 04, no. 01, pp. 11–23, 2015, doi: 10.17576/apjitm-2015-0401-02.

[21] M. Ter Hoeve, D. Grangier, and N. Schluter, "High-Resource Methodological Bias in Low-Resource Investigations," *ArXiv*, vol. abs/2211.07534, p., 2022, [Online]. Available: https://consensus.app/papers/highresource-methodological-bias-in-lowresource-hoeve-grangier/d1313c3c202b5621abea76a6fbb95aec/