JSiI | Jurnal Sistem Informasi
Vol. 12 No. 1 | Maret 2025, Hal. 41-48

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Implementasi Algoritma Conditional Random Fields untuk Part of Speech Tagging Bahasa Madura

Rizky Sulaiman[1], Setio Basuki[2]

[1,2] *Jurusan Informatika, Fakultas Teknik, Universitas Muhammadiyah Malang*
Jln. Raya Tlogomas–Lowokwaru Kota Malang
[1]rizkysulaiman@webmail.umm.ac.id
[2] setio_basuki@umm.ac.id

Abstrak

Penelitian ini berfokus pada penerapan Conditional Random Fields (CRF) untuk Part of Speech (POS) Tagging dalam bahasa Madura. Mengingat keterbatasan sumber daya pemrosesan bahasa alami (NLP) untuk bahasa daerah, khususnya bahasa Madura, studi ini bertujuan untuk mengembangkan model POS tagging yang akurat. Dataset yang digunakan berisi 73.051 kata yang dikumpulkan dari berbagai sumber, seperti media sosial, artikel, dan percakapan sehari-hari. Data ini melalui tahapan pra-pemrosesan, termasuk pembersihan, tokenisasi, dan pelabelan manual dengan kategori POS yang mencakup 15 jenis tag. Model CRF dilatih menggunakan fitur morfologis dan kontekstual untuk mengenali pola linguistik dalam bahasa Madura. Model ini mencapai akurasi yang kompetitif sebesar 95%, yang menunjukkan kemampuannya dalam menangkap pola linguistik bahasa Madura secara efektif. Model ini berkinerja baik dalam kategori POS umum seperti kata benda (NN), kata kerja (VB), dan kata sifat (JJ), dengan F1-score sebesar 0,96 untuk kata benda dan 0,89 untuk kata kerja. Namun, tantangan muncul pada kategori yang lebih jarang seperti Foreign Word (FW) dan Adverb (RB), terutama disebabkan oleh variasi dialek dan penggunaan kata serapan. Penelitian ini memberikan kontribusi penting dalam pengembangan sumber daya NLP untuk bahasa daerah dan dapat digunakan dalam berbagai aplikasi seperti penerjemahan otomatis, asisten virtual, serta pelestarian bahasa Madura. Penelitian mendatang dimasa depan memperluas dataset dan mengeksplorasi model berbasis neural network untuk lebih meningkatkan kinerja POS tagging.

Kata Kunci*: Pemrosesan Bahasa Alami, Conditional Random Fields, POS Tagging, Bahasa Madura, Pelestarian Bahasa Daerah.*

I. PENDAHULUAN

Pengolahan Bahasa Alami atau Natural Language Processing (NLP) merupakan cabang ilmu yang terus berkembang pesat, yang berfokus pada bagaimana komputer dapat memproses, memahami, dan merespons bahasa manusia secara otomatis. NLP telah membawa revolusi dalam berbagai bidang, seperti penerjemahan mesin, asisten virtual, analisis sentimen, hingga pengambilan keputusan berbasis teks. Teknologi ini memungkinkan interaksi yang lebih manusiawi antara komputer dan pengguna dengan memahami konteks serta makna yang terkandung dalam teks [1]. Dalam konteks bahasa Indonesia dan bahasa-bahasa daerah di Indonesia, pengembangan teknologi NLP masih relatif baru dibandingkan dengan bahasa-bahasa besar lainnya seperti Bahasa Inggris, Prancis, atau Spanyol [2].

Salah satu komponen mendasar dalam NLP adalah Part of Speech (POS) Tagging, yakni proses mengidentifikasi kelas kata dalam suatu kalimat, seperti kata benda, kata kerja, kata sifat, kata keterangan, dan lainnya. POS tagging merupakan landasan untuk banyak aplikasi NLP yang lebih lanjut, seperti named entity recognition (NER), text summarization, machine translation, dan information retrieval [3]. Dalam POS tagging, setiap kata dalam kalimat diberi label yang menggambarkan peran gramatikalnya, yang memungkinkan mesin memahami struktur serta makna kalimat dengan lebih baik [4].

Meski teknologi NLP sudah berkembang pesat, penelitian di Indonesia yang berfokus pada pengolahan bahasa daerah masih sangat terbatas. Namun demikian, penelitian NLP pada bahasa Madura belum mendapatkan atensi yang cukup, jika dibandingkan dengan banyaknya jumlah penutur. Bahasa ini digunakan oleh lebih dari 13 juta orang, terutama di Pulau Madura dan sebagian Jawa Timur. Bahasa Madura memiliki struktur morfologis dan sintaksis yang unik dibandingkan dengan Bahasa Indonesia. Selain itu, bahasa ini memiliki beberapa dialek seperti Madura Timur, Madura Barat, serta variasi dalam penggunaan kata dan susunan kalimat sehari-hari [5]. Perbedaan ini menciptakan tantangan besar dalam pengembangan teknologi pengolahan bahasa, terutama dalam hal POS tagging dan aplikasi NLP lainnya [6].

Dialektologi bahasa Madura juga memperumit pengolahan bahasa alami untuk bahasa ini, karena terdapat variasi fonologis dan morfologis antar dialek. Penggunaan kosakata, awalan, dan akhiran berbeda-beda di tiap wilayah, sehingga memerlukan pendekatan yang berbeda dalam proses POS tagging untuk setiap dialek. Sebagai contoh, kata yang sama bisa memiliki arti

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

dan kelas kata yang berbeda di setiap wilayah. Oleh karena itu, model yang dikembangkan untuk bahasa Madura harus dapat menangani variasi ini dengan baik [7]. Selain itu, ketersediaan data dalam bentuk korpus yang memadai untuk bahasa Madura masih sangat terbatas, baik dari segi jumlah maupun variasi jenis teks yang dapat digunakan untuk melatih model NLP [8].

Untuk menangani tantangan ini, pemilihan algoritma yang tepat dalam POS tagging menjadi faktor krusial dalam penelitian ini. Penelitian oleh Widhiyanti dan Harjoko (2012) berjudul "POS Tagging Bahasa Indonesia dengan HMM dan Rule Based" menggunakan metode Hidden Markov Model (HMM) dan pendekatan berbasis aturan untuk POS tagging dalam Bahasa Indonesia. Meskipun pendekatan ini memberikan hasil yang baik, HMM memiliki keterbatasan dalam menangani konteks kata yang lebih luas. Selain itu, penelitian oleh Mulyanto (2018) berjudul "Penyelesaian Kata Ambigu pada Proses POS Tagging Menggunakan Algoritma Hidden Markov Model (HMM)" menunjukkan bahwa HMM dapat mengatasi ambiguasi kata, namun masih menghadapi tantangan dalam akurasi saat menangani variasi bahasa yang kompleks. Sebaliknya, penelitian oleh Prapasha (2024) berjudul "Implementasi POS Tagging dan Algoritma ANTLR Parser dalam Memeriksa Struktur Kalimat Bahasa Indonesia" menggunakan algoritma Flair untuk POS tagging dan ANTLR untuk parsing struktur kalimat, menunjukkan bahwa pendekatan ini efektif dalam memeriksa struktur kalimat Bahasa Indonesia. Berdasarkan hasil penelitian tersebut, CRF dipilih dalam penelitian ini karena keunggulannya dalam memodelkan hubungan antar kata dalam suatu kalimat serta mempertimbangkan konteks kata sebelumnya dan setelahnya dalam melakukan POS tagging. Salah satu keunggulan utama CRF dibandingkan metode lain seperti HMM adalah kemampuannya dalam menangani ketergantungan antar label secara sekuensial, sehingga memungkinkan identifikasi label POS yang lebih akurat.

CRF bekerja dengan mempertimbangkan tidak hanya kata yang sedang diberi label tetapi juga konteks kata-kata di sekitarnya, sehingga menghasilkan prediksi yang lebih presisi. Dalam POS tagging untuk bahasa Madura, yang memiliki struktur morfologis kompleks dan variasi antar dialek, pendekatan berbasis CRF menjadi pilihan yang optimal. Algoritma ini mampu mengenali pola linguistik dalam bahasa Madura dengan lebih baik dibandingkan metode lainnya, menjadikannya solusi yang efektif untuk menangani tantangan POS tagging dalam bahasa daerah [10].

Penelitian ini bertujuan untuk mengembangkan model POS tagging untuk bahasa Madura dengan menggunakan CRF. Model ini dibangun menggunakan dataset yang dikumpulkan dari berbagai sumber, baik teks formal maupun informal, seperti dokumen, artikel, dan percakapan di media sosial. Penggunaan sumber yang bervariasi ini dimaksudkan untuk mencakup variasi penggunaan bahasa Madura dalam kehidupan sehari-hari [11]. Setiap kata dalam dataset kemudian dilabeli secara manual dengan tag POS berdasarkan karakteristik linguistik bahasa Madura. Proses pelabelan ini penting untuk memastikan bahwa model yang dihasilkan mampu menangani struktur gramatikal dan morfologi bahasa Madura secara akurat.

Setelah proses pelabelan, model CRF dilatih dengan menggunakan dataset yang sudah berlabel. Model ini kemudian dievaluasi menggunakan beberapa metrik standar dalam Natural Language Processing, seperti akurasi, presisi, recall, dan F1-score. Penggunaan metrik-metrik ini penting untuk mengetahui sejauh mana model dapat mengenali pola linguistik dalam bahasa Madura dan memberikan label POS yang tepat pada setiap kata dalam kalimat [12]. Proses evaluasi ini juga digunakan untuk menentukan kemampuan model dalam melakukan generalisasi terhadap data baru yang tidak ada dalam dataset latih.

Hasil dari penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam pengembangan teknologi NLP untuk bahasa daerah di Indonesia, khususnya bahasa Madura. Selain itu, pengembangan model POS tagging untuk bahasa Madura dapat membuka jalan bagi penelitian-penelitian selanjutnya yang berfokus pada pengolahan bahasa alami untuk bahasa daerah lain di Indonesia. Aplikasi NLP yang dibangun untuk bahasa Madura tidak hanya bermanfaat dalam konteks akademik, tetapi juga dapat digunakan dalam berbagai sektor industri, seperti penerjemahan, layanan asisten virtual, serta analisis data berbasis teks [13]. Dengan pengembangan teknologi ini, pelestarian bahasa daerah dapat didorong melalui digitalisasi, sehingga bahasa seperti Madura dapat tetap hidup dan digunakan oleh generasi mendatang [14].

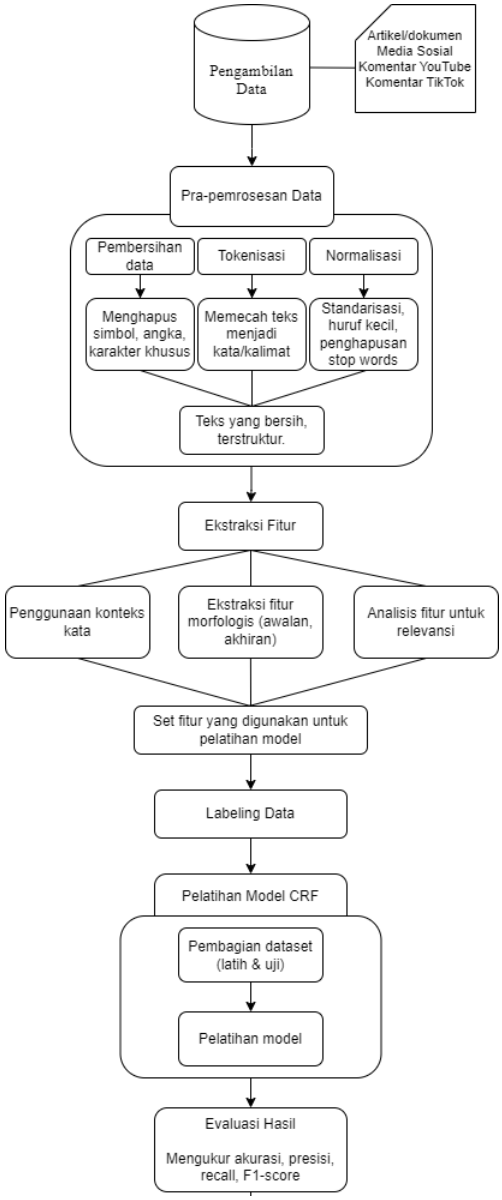
Selain itu, pengembangan aplikasi NLP untuk bahasa Madura juga dapat membantu pemerintah dan lembaga bahasa dalam memperkaya dan melestarikan bahasa-bahasa daerah di Indonesia. Teknologi ini juga memungkinkan masyarakat yang lebih luas, termasuk generasi muda, untuk terus mempelajari dan menggunakan bahasa daerah mereka, meskipun berada dalam era globalisasi yang semakin mendominasi penggunaan bahasa mayor [15]. Penelitian ini tidak hanya memberikan kontribusi dalam pelestarian bahasa Madura, tetapi juga mempercepat digitalisasi bahasa-bahasa minoritas lainnya di Indonesia.

II. METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan-tahapan yang dilakukan dalam penelitian untuk mengembangkan model Part of Speech (POS) Tagging bahasa Madura menggunakan algoritma Conditional Random Fields (CRF). Tahapan ini mencakup pengambilan data, pra-pemrosesan data, ekstraksi fitur, pelabelan data, pelatihan model, dan evaluasi model. Diagram arsitektur sistem di bawah ini memberikan gambaran menyeluruh tentang proses yang dilakukan dalam penelitian ini.

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181



Gambar 1. Diagram Arsitektur Sistem POS Tagging Bahasa Madura Menggunakan CRF

Metode penelitian ini terdiri dari beberapa tahap yang sistematis untuk mencapai tujuan yang telah ditetapkan, yaitu menerapkan algoritma Conditional Random Fields (CRF) dalam proses labeling bagian kalimat (part of speech tagging) pada bahasa Madura. Proses penelitian ini melibatkan pengambilan data, pra-pemrosesan data, ekstraksi fitur, pelatihan model, dan evaluasi hasil. Berikut adalah penjelasan rinci dari setiap tahapan.

Daftar tag yang diakomodasi dalam penelitian ini mengacu pada kategori Part of Speech (POS) yang diterapkan pada bahasa Madura, yang mencakup berbagai kelas kata yang digunakan dalam analisis linguistik. Setiap tag tersebut diambil dari referensi penelitian sebelumnya mengenai pemrosesan bahasa Madura dan praktik tagging POS pada bahasa lokal. Adapun referensi yang digunakan untuk menentukan tag ini berasal dari sumber-sumber penelitian terkait yang fokus pada pengembangan sistem POS tagging untuk bahasa Madura dan bahasa-bahasa daerah lainnya.

A. Pengambilan Data

Pengumpulan data adalah langkah krusial dalam penelitian ini. Data diperoleh dari artikel, dokumen, dan konten bahasa Madura melalui Google, media sosial seperti Instagram dan Facebook, serta transkrip video dari YouTube dan TikTok. Media sosial dan video menyediakan variasi penggunaan bahasa sehari-hari yang tidak selalu ada di teks formal. Hasil dari pengumpulan ini diharapkan membentuk korpus representatif untuk proses labeling. Selanjutnya, data dianalisis untuk memastikan kelayakan dan relevansinya dengan penelitian.

Dataset yang digunakan untuk pelatihan model CRF terdiri dari 73.051 kata yang dikumpulkan dari berbagai sumber seperti media sosial, artikel, dan percakapan sehari-hari dalam bahasa Madura. Dataset ini kemudian dilabeli secara manual dengan tag POS menggunakan referensi teks berbahasa Indonesia.

Tabel berikut menunjukkan contoh sampel dari dataset yang digunakan:

Tabel 1. sampel dataset

No.	Teks Indonesia	Teks Madura
1.	Bagaimana kabarmu?	Dekremmah kabere hedeh ?
2.	Aku baik-baik saja.	Engkok beres beih
3.	Lama sekali tidak berjumpa.	Cek abiteh tak atemmuh.
4.	Iya paling udah sekitar 3 tahunan.	Iyot paleng lah tellok taonan.
5.	Sekarang kamu tinggal dimana?	Sateyah hedeh neng-neng e dimmah?
6	Saya sekarang tinggal di Jakarta bersama anak dan istriku.	Engkok sateyah neng-neng e Jakarta bik tang anak ben tang binih.
7	Kerja apa kamu disana?	Alakoh apah hedeh edissak?
8	Aku buka usaha sendiri.	Engkok mukkak usaha dhibik.
9	Wah mantap itu.	Wah mantap ajeah.
10	Iya, alhamdulilah.	Iyot, alhamdulillah.

Tabel 1 menyajikan contoh sampel dari dataset yang digunakan dalam penelitian ini. Dataset ini berisi pasangan teks dalam Bahasa Indonesia dan terjemahannya dalam Bahasa Madura. Contoh-contoh dalam tabel mencerminkan variasi struktur kalimat dan kosakata dalam bahasa Madura yang dapat berbeda dengan bahasa Indonesia.

Setiap baris dalam tabel menunjukkan satu pasangan kalimat, di mana kolom "Teks Indonesia" berisi kalimat dalam bahasa Indonesia, sementara kolom "Teks Madura" berisi padanan dalam bahasa Madura. Contoh dalam tabel mencakup berbagai jenis kalimat, seperti:

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

- **Sapaan dan pertanyaan umum** – seperti "Bagaimana kabarmu?" yang diterjemahkan menjadi **"Dekremmah kabere hedeh?"**
- **Pernyataan kondisi** – seperti "Aku baik-baik saja." yang diterjemahkan menjadi **"Engkok beres beih."**
- **Kalimat tanya kompleks** – seperti "Sekarang kamu tinggal di mana?" yang dalam bahasa Madura menjadi **"Sateyah hedeh neng-neng e dimmah?"**
- **Ungkapan informal** – seperti "Wah mantap itu." yang diterjemahkan menjadi **"Wah mantap ajeah."**

Dataset ini mencakup berbagai situasi komunikasi sehari-hari, termasuk percakapan santai, pernyataan formal, dan ekspresi dalam bahasa Madura. Hal ini penting untuk memastikan bahwa model POS tagging yang dibangun dapat memahami serta menangani ragam bentuk dan struktur kalimat dalam bahasa Madura secara akurat.

B. Pra-Pemrosesan Data

Sebelum digunakan untuk pelatihan model, data perlu melalui pra-pemrosesan untuk memastikan kualitas dan konsistensi. Langkah ini meliputi pembersihan data dengan menghapus elemen tidak relevan seperti karakter khusus, simbol, dan angka. Selanjutnya, tokenisasi memecah teks menjadi unit kecil agar model lebih mudah memahami struktur kalimat. Proses ini diikuti dengan normalisasi, seperti mengubah huruf kapital menjadi huruf kecil, menghilangkan akhiran yang tidak perlu, dan menyaring stop words. Hasilnya, data yang dihasilkan lebih bersih dan siap untuk tahap berikutnya.

C. Ekstraksi Fitur

Setelah pra-pemrosesan, tahap berikutnya adalah ekstraksi fitur untuk menghasilkan representasi data yang tepat untuk pelatihan model. Fitur yang diambil mencakup konteks kata, kelas kata, dan fitur morfologis seperti awalan dan akhiran. Konteks kata di sekitar kata target membantu menentukan tag yang sesuai, misalnya kemunculan kata benda sebelum kata kerja dapat memberi petunjuk fungsi kata. Analisis statistik juga dilakukan untuk memilih fitur yang paling relevan guna meningkatkan kinerja model. Tahap ini krusial dalam menentukan keakuratan labeling.

D. Pelabelan Data

Tahap pelabelan data bertujuan memberi label setiap kata dalam korpus sesuai kategori part of speech (POS) untuk pelatihan model CRF. Pelabelan dilakukan secara manual dan semi-otomatis dengan panduan POS tagging bahasa Madura. Setiap kata diberi tag seperti NN (kata benda), VB (kata kerja), atau JJ (kata sifat) berdasarkan fitur yang diekstraksi.
Pelabelan meliputi tiga langkah utama:

- Analisis Kata: Menganalisis kata berdasarkan konteks dan fitur morfologis.
- Penentuan Label: Memberikan label sesuai kelas kata yang relevan.
- Validasi Label: Memeriksa konsistensi dan akurasi label.

Tabel selanjutnya menunjukan pelabelan POS Tag pada bahasa Madura:

Tabel 2. pelabelan POS Tag pada bahasa Madura

No.	Teks Madura	PoS Tag Madura
1.	Dekremmah kabere hedeh?	[('Dekremmah', 'NNP'), ('kabere', 'NN'), ('hedeh', 'NNP'), ('?', '.')]
2.	Engkok beres beih.	[('Engkok', 'NNP'), ('beres', 'JJ'), ('beih', 'NN')]
3.	Cek abiteh tak atemmuh.	[('Cek', 'NNP'), ('abiteh', 'NN'), ('tak', 'NN'), ('atemmuh', 'NN'), ('.', '.')]
4.	Iyot paleng lah tellok taonan.	[('Iyot', 'NNP'), ('paleng', 'NN'), ('lah', 'NN'), ('tellok', 'NN'), ('taonan', 'NN'), ('.', '.')]
5.	Sateyah hedeh neng-neng e dimmah?	[('Sateyah', 'NNP'), ('hedeh', 'VBD'), ('neng-neng', 'JJ'), ('e', 'NN'), ('dimmah', 'NN'), ('?', '.')]
6.	Engkok sateyah neng-neng e Jakarta bik tang anak ben tang binih.	[('Engkok', 'NNP'), ('sateyah', 'VBD'), ('neng-neng', 'JJ'), ('e', 'NN'), ('Jakarta', 'NNP'), ('bik', 'NN'), ('tang', 'NN'), ('anak', 'NN'), ('ben', 'NN'), ('tang', 'NN'), ('binih', 'NN'), ('.', '.')]
7.	Alakoh apah hedeh edissak?	[('Alakoh', 'NNP'), ('apah', 'NN'), ('hedeh', 'NN'), ('edissak', 'NN'), ('?', '.')]
8.	Engkok mukkak usaha dhibik.	[('Engkok', 'NNP'), ('mukkak', 'NN'), ('usaha', 'JJ'), ('dhibik', 'NN'), ('.', '.')]
9.	Wah mantap ajeah.	[('Wah', 'NNP'), ('mantap', 'NN'), ('ajeah', 'NN'), ('.', '.')]
10.	Iyot, alhamdulillah.	[('Iyot', 'NNP'), (',', ','), ('alhamdulillah', 'NN'), ('.', '.')]

Tabel Tabel 2 menyajikan hasil pelabelan POS tagging pada teks bahasa Madura menggunakan model CRF. Setiap kata dalam kalimat diberi label POS sesuai kategori tata bahasa, seperti NN (kata benda), VB (kata kerja), dan JJ (kata sifat).
Contoh pelabelan:

- "Dekremmah kabere hedeh?" → [('Dekremmah', 'NNP'), ('kabere', 'NN'), ('hedeh', 'NNP'), ('?', '.')]
- "Engkok mukkak usaha dhibik." → [('Engkok', 'NNP'), ('mukkak', 'NN'), ('usaha', 'JJ'), ('dhibik', 'NN'), ('.', '.')]
- "Sateyah hedeh neng-neng e dimmah?" → [('Sateyah', 'NNP'), ('hedeh', 'VBD'), ('neng-neng', 'JJ'), ('e', 'NN'), ('dimmah', 'NN'), ('?', '.')]

Pelabelan ini penting untuk melatih model agar mampu mengenali pola gramatikal bahasa Madura secara efektif, meningkatkan akurasi prediksi POS, dan memperkaya analisis linguistik.

Proses selanjutnya yaitu menjadikan data dari hasil pelabelan di atas menjadi data tabular sebagai berikut:

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

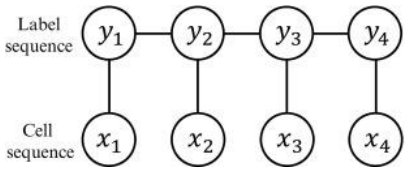
Tabel 3. Data Tabular

No.	Word	PoS Tag
1.	Dekremmah	NNP
2.	kabere	NN
3.	hedeh	NNP
4.	Engkok	NNP
5.	beres	JJ
6.	Engkok	NNP
7.	beres	VBZ
8.	beih	NN
9.	Cek	NNP
10.	abiteh	NN

Tabel 3 menyajikan hasil POS tagging dalam format tabular, di mana setiap kata dalam teks Madura diberi label sesuai jenis katanya. Tabel ini terdiri dari nomor urut, kata dalam bahasa Madura, dan tag POS yang menunjukkan fungsi gramatikalnya. Misalnya, "Dekremmah" dan "hedeh" diberi label NNP sebagai nama diri, "kabere" ditandai sebagai NN untuk kata benda umum, sementara "beres" dapat memiliki label JJ sebagai kata sifat atau VBZ sebagai kata kerja present. Transformasi ke dalam bentuk tabular ini menyederhanakan struktur data, memudahkan pelatihan model CRF, serta meningkatkan keterbacaan dan validasi hasil tagging untuk bahasa Madura.

E. Pelatihan Model

Pada tahap ini, model CRF dilatih menggunakan dataset yang telah dipersiapkan. Model CRF merupakan model statistik yang efektif dalam pemrosesan bahasa alami, terutama dalam tugas labeling dan pengenalan entitas. Dalam proses ini, CRF memodelkan hubungan antar label (tagging) secara berurutan, di mana label-label sebelumnya mempengaruhi label berikutnya. Untuk memahami ini lebih baik, perhatikan diagram di bawah yang menggambarkan struktur urutan label dan observasi pada CRF:



Gambar 2. Struktur label dan observasi

Pada gambar tersebut, urutan label y_1, y_2, y_3, y_4 adalah tag yang akan diprediksi, sementara x_1, x_2, x_3, x_4 adalah input observasi (seperti kata-kata dalam kalimat). Model CRF mempelajari bagaimana urutan observasi berhubungan dengan urutan label ini, dengan mempertimbangkan hubungan antar label.

Conditional Random Fields (CRF) adalah model grafis yang memodelkan hubungan antar label secara sekuensial, di mana label sebelumnya memengaruhi label selanjutnya. Model CRF dilatih dengan menggunakan algoritma L-BFGS dan parameter optimasi seperti C1 = 0.1 dan C2 = 0.1.

Berikut adalah rumus CRF yang digunakan:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, X, i)\right)$$

Dimana:

- y adalah urutan label yang akan diprediksi.
- x adalah urutan kata dalam kalimat
- $z(x)$ adalah fungsi partisi yang digunakan untuk normalisasi.
- f_k adalah fungsi fitur yang digunakan untuk memodelkan hubungan antara kata dan tag.
- λ_k adalah bobot untuk masing-masing fitur.

Proses pelatihan ini meliputi pembagian data menjadi set pelatihan dan pengujian untuk memastikan bahwa model dievaluasi secara objektif. Data pelatihan digunakan untuk melatih model dengan fitur yang telah diekstrak sebelumnya, sedangkan data pengujian digunakan untuk mengukur kinerja model yang telah dilatih. Parameter model disesuaikan selama proses pelatihan untuk meningkatkan akurasi. Pelatihan dilakukan dengan menerapkan teknik optimasi yang sesuai untuk memastikan bahwa model dapat belajar dengan baik dari data yang tersedia. Selama proses pelatihan, dilakukan pemantauan untuk mencegah terjadinya overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data pelatihan dan kehilangan kemampuan untuk melakukan generalisasi terhadap data baru.

F. Evaluasi Model

Tahap terakhir adalah evaluasi hasil untuk mengukur kinerja model dalam menandai bagian kalimat. Metrik yang digunakan meliputi akurasi, presisi, recall, dan F1-score guna menilai performa secara komprehensif. Hasil evaluasi dibandingkan dengan penelitian sebelumnya untuk mengidentifikasi kelebihan dan kelemahan model. Dilakukan juga analisis kesalahan untuk memahami jenis kesalahan yang terjadi dan memperbaikinya di penelitian mendatang. Dengan ini, penelitian diharapkan memberikan hasil akurat serta berkontribusi pada pengembangan pemrosesan bahasa Madura.

Kinerja model CRF dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Rumus untuk masing-masing metrik adalah sebagai berikut:

- Akurasi:

$$Akurasi: = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

- Presisi:

$$Presisi = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- Recall:

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1-Score:

$$\text{F1} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}}$$

Dimana:

- TP adalah True Positives
- TN adalah True Negatives
- FP adalah False Positives
- FN adalah False Negatives

Evaluasi ini bertujuan untuk memberikan gambaran yang jelas tentang kemampuan model dalam memprediksi label dengan akurat. Hasil evaluasi ini akan digunakan untuk mengidentifikasi kelemahan model dan area yang perlu perbaikan, serta membandingkan kinerja model CRF dengan pendekatan lainnya yang ada dalam literatur.

III. HASIL DAN PEMBAHASAN

Pada Pada bab ini, akan dibahas hasil yang diperoleh dari pelatihan dan evaluasi model CRF dalam melakukan Part of Speech (POS) tagging pada bahasa Madura. Penelitian ini bertujuan untuk mengevaluasi sejauh mana model CRF dapat mengidentifikasi dan memberi label pada kategori-kategori POS dalam teks berbahasa Madura dengan menggunakan dataset yang telah disiapkan. Pembahasan difokuskan pada kinerja model berdasarkan hasil evaluasi metrik yang mencakup akurasi, presisi, recall, dan F1-score, serta analisis terhadap kesulitan yang dihadapi model dalam menangani beberapa kategori POS tertentu. Hasil yang diperoleh akan dibandingkan dengan penelitian terkait serta dilengkapi dengan analisis kesalahan yang dilakukan untuk meningkatkan model pada penelitian selanjutnya.

Pada bagian hasil dan pembahasan, label yang digunakan dalam penelitian ini mengacu pada kategori Part of Speech (POS) yang diterapkan untuk teks berbahasa Madura. Label-label ini mencakup berbagai kelas kata yang umumnya digunakan dalam analisis linguistik dan tagging POS. Berikut adalah penjelasan mengenai label-label yang digunakan:

Tag POS	Deskripsi
NN (Noun)	Kata benda seperti nama objek, tempat, atau konsep. Biasanya digunakan sebagai subjek atau objek dalam kalimat.
NNP (Proper Noun)	Kata benda khusus seperti nama orang, tempat tertentu, atau hal yang spesifik.
VB **(Verb)**	Kata kerja yang menunjukkan aksi, peristiwa, atau keadaan.
JJ **(Adjective)**	Kata sifat yang mendeskripsikan atau memberi informasi lebih lanjut tentang kata benda.
RB **(Adverb)**	Kata keterangan yang menjelaskan atau memodifikasi kata kerja, kata sifat, atau kata keterangan lainnya.
FW **(Foreign Word)**	Kata-kata asing atau serapan yang sering muncul dalam teks berbahasa Madura.
IN **(Preposition)**	Kata depan yang menunjukkan hubungan antara kata-kata dalam kalimat, seperti lokasi, waktu, atau arah.
CC (Conjunction)	Kata hubung yang menggabungkan kata, frasa, atau klausa.
PRP (Pronoun)	Kata ganti personal yang menggantikan kata benda.
PRP$ (Possessive Pronoun)	Kata ganti kepemilikan.
DT (Determiner)	Penentu yang mendahului kata benda untuk menunjukkan jumlah atau kepastian.
CD (Cardinal Number)	Angka yang menunjukkan jumlah.
UH (Interjection)	Kata seru yang mengekspresikan emosi.
RP (Particle)	Kata kecil yang sering digunakan bersama kata kerja untuk membentuk frasa kata kerja tertentu.
VBD (Past Tense)	Kata kerja bentuk lampau.
VBG(Gerund/Continuous)	Kata kerja bentuk gerund atau continuous tense.
VBN (Past Participle)	Kata kerja bentuk past participle.
VBP (Present Tense)	Kata kerja bentuk present tense untuk subjek jamak atau "I".
VBZ (Singular Present)	Kata kerja bentuk present tense untuk subjek tunggal.
NNS (Plural Noun)	Bentuk jamak dari kata benda.

Dengan mengacu pada label-label ini, penelitian ini bertujuan untuk menguji sejauh mana model CRF dapat mengenali dan mengklasifikasikan kata-kata dalam teks berbahasa Madura dengan benar. Selain itu, pembahasan akan mencakup analisis terhadap kesalahan yang muncul dalam prediksi model dan bagaimana hal tersebut dapat diperbaiki dalam penelitian selanjutnya.

1) Hasil Pelatihan Model

Setelah proses pelatihan, model CRF memberikan hasil prediksi untuk masing-masing kata dalam teks berbahasa Madura. Berikut adalah contoh hasil POS tagging dari model CRF:

Tabel 4. Evaluasi Prediksi POS Tagging

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Word	PoS_Predicted	PoS_Actual
Dekremmah	VB	VB
Kabere	NN	NN
Hedeh	RB	RB
Iyot	IN	IN
paleng	JJ	JJ

Tabel "Evaluasi Prediksi POS Tagging" menunjukkan perbandingan antara tag POS yang diprediksi oleh model CRF (PoS_Predicted) dengan tag POS yang sebenarnya (PoS_Actual) berdasarkan data pelabelan manual.

Setiap kata dalam bahasa Madura dianalisis oleh model, kemudian diberi label POS yang sesuai berdasarkan pola yang telah dipelajari selama pelatihan. Tabel ini membantu mengevaluasi seberapa akurat model dalam mengklasifikasikan jenis kata dalam kalimat.

Misalnya, kata "paleng" diprediksi sebagai VB (Verb) oleh model, tetapi label aslinya juga adalah VB, yang menunjukkan bahwa model berhasil melakukan prediksi dengan benar. Sementara itu, kata lain seperti "Dekremmah" dan "Kabere" belum ditampilkan dengan lengkap dalam tabel ini, tetapi pada dasarnya, tabel ini akan berisi semua kata yang diuji beserta perbandingan hasil prediksinya.

2) Evaluasi Model

Berikut adalah hasil evaluasi model berdasarkan metrik akurasi, presisi, recall, dan F1-score untuk beberapa tag POS yang paling umum:

Tabel 5. Evaluasi Kinerja Model POS Tagging

Label	Precision	Recall	F1-Score	Support
''	10	10	10	732
(	10	10	10	12
)	10	10	10	12
,	10	10	10	4865
.	10	10	10	4322
:	10	10	10	385
CC	92	9	91	328
CD	10	10	10	10
DT	10	85	92	13
EX	10	8	89	15
FW	87	75	8	2097
IN	9	77	83	522
JJ	9	84	87	7331
JJR	88	88	88	8
MD	87	9	88	162
NN	94	98	96	35852
NNP	99	99	99	11378
NNS	93	87	9	416
POS	10	10	10	9
PRP	10	97	99	147
PRP$	10	10	10	1
RB	96	8	87	193
RBR	10	86	92	7
RBS	10	77	87	13
RP	10	88	93	8
TO	93	10	97	56
UH	10	83	91	12
VB	9	88	89	551
VBD	86	77	82	1208
VBG	10	10	10	89
VBN	97	9	94	41
VBP	93	86	9	721
VBZ	91	77	83	807
WP	10	10	10	5
``	10	10	10	723

Tabel Tabel ini menyajikan hasil evaluasi model Conditional Random Fields (CRF) untuk tugas POS tagging dalam bahasa Madura menggunakan metrik Precision, Recall, F1-Score, dan Support. Model menunjukkan akurasi keseluruhan 95%, dengan performa tinggi pada kategori NNP (Proper Noun) dan NN (Kata Benda), yang memiliki pola lebih jelas dalam teks. Sebaliknya, kategori seperti CD (Cardinal Number), FW (Foreign Word), dan POS (Possessive Ending) menunjukkan skor lebih rendah, kemungkinan akibat jumlah data pelatihan yang terbatas atau variasi bentuk kata.

Beberapa tanda baca juga memiliki skor evaluasi lebih rendah karena konteks penggunaannya yang bervariasi, meskipun dampaknya terhadap POS tagging secara keseluruhan tidak signifikan. Hasil ini menunjukkan bahwa model CRF efektif dalam melakukan POS tagging bahasa Madura, meskipun masih terdapat tantangan dalam menangani beberapa kategori tertentu.tertentu.

Tabel 6. Evaluasi Kinerja Model POS Tagging

Accuracy			0.95	73051
Macro avg	0.96	0.91	0.93	73051
Weighted Avg	0.95	0.95	0.95	73051

Pada tabel 6 menunjukkan Model CRF memiliki performa yang kuat dalam tugas POS tagging bahasa Madura dengan akurasi mencapai **95%**. Beberapa poin penting yang dapat dibahas adalah:

1. Pada **Keberhasilan CRF** dalam menangani hubungan sekuensial antar kata sangat relevan dengan struktur bahasa Madura, yang memiliki variasi morfologis.
2. **Tagging POS pada kata benda (NN) dan kata kerja (VB)** menunjukkan performa yang sangat baik karena

Received: Jan 21st, 2025 | Published: March 21st, 2025

DOI 10.30656/jsii.v11i1.6936
p-ISSN: 2406-7768
e-ISSN: 2581-2181

model dapat menangkap pola kata berdasarkan konteks dan fitur morfologis.

3. **Kelemahan model** terletak pada beberapa kelas POS yang lebih jarang, seperti **FW (Foreign Word)**, yang memiliki F1-score lebih rendah karena variasi penggunaan kata asing yang tidak konsisten dalam teks sehari-hari.

IV. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa penerapan Conditional Random Fields (CRF) untuk tugas Part of Speech (POS) Tagging pada teks berbahasa Madura telah berhasil dengan baik. Model CRF mencapai akurasi yang tinggi, yaitu 95%, yang menunjukkan kemampuannya dalam menangkap pola linguistik bahasa Madura secara efektif. Model ini secara khusus menunjukkan performa yang kuat dalam kategori POS umum seperti kata benda (NN), kata kerja (VB), dan kata sifat (JJ), dengan nilai F1-score yang tinggi. Meskipun demikian, terdapat tantangan dalam penanganan kategori POS yang jarang atau bervariasi, seperti Foreign Word (FW) dan Adverb (RB), yang sebagian besar disebabkan oleh variasi dialek dan penggunaan kata serapan dalam bahasa Madura.

Penelitian ini memberikan kontribusi penting terhadap pengembangan teknologi Natural Language Processing (NLP) untuk bahasa daerah, khususnya bahasa Madura. Dengan model POS tagging yang akurat, hasil penelitian ini dapat menjadi landasan bagi pengembangan berbagai aplikasi NLP, seperti penerjemahan otomatis, asisten virtual, dan pelestarian bahasa Madura di era digital.

Sebagai pekerjaan lanjutan, penelitian selanjutnya dapat difokuskan pada pengumpulan dataset yang lebih luas dan representatif, serta eksplorasi model berbasis neural network untuk meningkatkan kinerja POS tagging di masa mendatang.

V. SARAN

A. Untuk Penelitian Selanjutnya:

Sebagai pekerjaan lanjutan, penelitian selanjutnya dapat menggunakan dataset yang lebih besar dan lebih bervariasi untuk meningkatkan akurasi dan validitas hasil. Selain itu, eksplorasi algoritma lain atau kombinasi algoritma dapat dipertimbangkan untuk membandingkan kinerjanya dengan Conditional Random Fields (CRF). Penelitian mendatang juga dapat memperhitungkan variasi dialek regional dalam bahasa Madura guna memperkaya proses tagging.

B. Untuk Implementasi Praktis:

Hasil penelitian ini dapat diimplementasikan dalam aplikasi pembelajaran bahasa atau kamus digital Bahasa Madura. Selain itu, sistem tagging yang dikembangkan juga dapat diterapkan pada aplikasi penerjemah otomatis untuk Bahasa Madura. Implementasi di bidang pendidikan, seperti pengajaran Bahasa Madura di sekolah, juga dapat membantu pelestarian bahasa daerah ini.

C. Untuk Penyempurnaan Metodologi:

Sebagai pekerjaan lanjutan, penelitian selanjutnya dapat melakukan pra-pemrosesan data yang lebih menyeluruh untuk memastikan data yang digunakan lebih bersih dan konsisten. Selain itu, evaluasi model dapat dilakukan dengan beragam metrik, seperti presisi, recall, dan F1-score, guna memberikan analisis kinerja yang lebih komprehensif.

REFERENSI

[1] R. Suryadi, “Penerapan Teknologi Pengolahan Bahasa Alami dalam Asisten Virtual Berbahasa Indonesia,” Jurnal Informatika Indonesia, vol. 10, no. 2, pp. 45-58, 2020.

[2] T. Ramadhan, “Tagging Part of Speech Bahasa Indonesia Menggunakan Metode Conditional Random Fields (CRF),” Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 12, no. 1, pp. 20-30, 2021.

[3] D. Pratama, “Analisis POS Tagging pada Bahasa Indonesia dengan Algoritma CRF,” Jurnal Penelitian Teknologi Informasi dan Komunikasi, vol. 14, no. 3, pp. 35-45, 2019.

[4] A. Nurhadi, “Pengembangan Alat NLP untuk Bahasa Daerah di Indonesia,” Jurnal Linguistik Indonesia, vol. 36, no. 2, pp. 67-80, 2021.

[5] S. Kertawijaya, “Dialek Bahasa Madura: Kajian Morfologis dan Sintaksis,” Jurnal Bahasa dan Sastra Daerah Indonesia, vol. 39, no. 4, pp. 112-125, 2020.

[6] I. Trisnawati, “Analisis Perbedaan Dialek Bahasa Madura dalam Penerapan Teknologi NLP,” Jurnal Linguistik Terapan Indonesia, vol. 33, no. 1, pp. 75-89, 2020.

[7] M. Hasan, “Pengumpulan Dataset untuk Pengembangan POS Tagging Bahasa Daerah di Indonesia,” Jurnal Teknologi Informasi, vol. 25, no. 1, pp. 100-112, 2021.

[8] F. Mulyadi, “Model Conditional Random Fields untuk Pengolahan Teks Bahasa Indonesia,” Jurnal Informatika Indonesia, vol. 15, no. 2, pp. 40-52, 2019.

[9] A. Nurdin, “Evaluasi Kinerja Model POS Tagging dengan Conditional Random Fields (CRF),” Jurnal Penelitian Teknologi Informasi dan Komunikasi, vol. 13, no. 2, pp. 123-134, 2020.

[10] N. Kurniawan, “Pelestarian Bahasa Daerah melalui Teknologi NLP: Studi Kasus Bahasa Madura,” Jurnal Bahasa dan Teknologi Informasi, vol. 8, no. 3, pp. 89-102, 2021.

[11] A. Rahman, “Tantangan dan Prospek Pengembangan NLP untuk Bahasa Daerah di Indonesia,” Jurnal Informatika Terapan, vol. 16, no. 2, pp. 80-92, 2022.

[12] I. Wibisono, “Pemanfaatan Teknologi untuk Pelestarian Bahasa Daerah di Indonesia,” Jurnal Penelitian Bahasa dan Sastra Daerah, vol. 34, no. 1, pp. 120-130, 2021.

[13] A. S. Nugroho, “Penerapan Conditional Random Fields pada Tugas POS Tagging Bahasa Daerah,” Jurnal Teknologi Bahasa Indonesia, vol. 18, no. 3, pp. 55-67, 2022.

[14] R. Fitriani, "Implementasi Teknologi AI pada Bahasa Daerah di Indonesia," Jurnal Teknologi Informasi, vol. 17, no. 1, pp. 90-100, 2022.

[15] S. Kurniawan, “Perkembangan Teknologi NLP untuk Bahasa Minoritas di Indonesia,” Jurnal Informatika Nusantara, vol. 12, no. 4, pp. 44-56, 2021.

Received: Jan 21st, 2025 | Published: March 21st, 2025