NAMED ENTITY RECOGNITION PADA TEKS BERBAHASA INDONESIA MENGGUNAKAN CNNs

Alfin Gunawan [1], M. Wanda[2], Rini Meiyanti[3]

Program Studi Teknik Informatika, Universitas Malikussaleh, Lhokseumawe, Aceh
Email: [1] alfin.210170021@mhs.unimal.ac.id , [2] wanda.210170041@mhs.unimal.ac.id , [3] rinimeiyanti@unimal.ac.id .

Abstrak

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model *Named Entity Recognition* (NER) pada teks berbahasa Indonesia menggunakan arsitektur *Convolutional Neural Networks* (CNNs). Proses penelitian dimulai dengan pengumpulan *dataset* yang melibatkan artikel berita, dokumen formal, dan transkripsi suara ke teks, menggunakan teknik *web scraping* untuk mendapatkan data secara otomatis. *Dataset* ini kemudian melalui proses *preprocessing* yang mencakup *case folding* dan *tokenization*. Selanjutnya, data diberi label menggunakan format *Stanford* BIO untuk mengidentifikasi entitas. Pembagian data dilakukan menjadi set pelatihan, pengujian, dan validasi dengan berbagai skenario. Model dilatih dan dievaluasi berdasarkan metrik presisi, *recall*, dan *F1-score* untuk mengukur kinerja dalam mengidentifikasi entitas seperti nama orang, organisasi, dan lokasi. Hasil penelitian menunjukkan bahwa pendekatan CNNs efektif dalam meningkatkan akurasi pengenalan entitas dalam teks berbahasa Indonesia, dengan F1 *Score* yang tinggi untuk kategori entitas yang berbeda, menunjukkan potensi signifikan untuk aplikasi lebih lanjut dalam pengolahan bahasa alami di Indonesia.

Kata Kunci: *Named Entity Recognition, Convolutional Neural Networks*

Abstract

This research aims to develop and evaluate a Named Entity Recognition (NER) model on Indonesian text using Convolutional Neural Networks (CNNs) architecture. The research process begins with the collection of datasets involving news articles, formal documents, and voice-to-text transcriptions, using web scraping techniques to obtain data automatically. This dataset then goes through a preprocessing process that includes case folding and tokenization. Next, the data is labeled using the Stanford BIO format to identify entities. Data division was done into training, testing, and validation sets with various scenarios. Models were trained and evaluated based on precision, recall, and F1-score metrics to measure performance in identifying entities such as names of people, organizations, and locations. The results show that the CNNs approach is effective in improving entity recognition accuracy in Indonesian text, with a high F1 Score for different entity categories, indicating significant potential for further applications in natural language processing in Indonesia.

Keywords: *Named Entity Recognition, Convolutional Neural Networks*

---

1. PENDAHULUAN

Named Entity Recognition (NER) merupakan salah satu komponen kunci dalam bidang pengolahan bahasa alami atau *Natural Language Processing* (NLP) [1]. Fungsi utamanya adalah untuk mengidentifikasi dan mengklasifikasikan entitas penting dalam teks, seperti nama orang, organisasi, lokasi, dan tanggal, ke dalam kategori-kategori yang sudah ditentukan sebelumnya [2]. Kemampuan NER untuk menyaring informasi ini sangat penting dalam berbagai aplikasi seperti sistem pencarian informasi, terjemahan mesin, analisis opini, dan asisten virtual. Dalam konteks bahasa Indonesia, pengembangan NER menghadapi tantangan-tantangan tersendiri yang berbeda dengan bahasa-bahasa lain yang lebih banyak digunakan, seperti bahasa Inggris atau Mandarin. Salah satu hambatan utamanya adalah kurangnya ketersediaan data yang memadai serta struktur bahasa Indonesia yang memiliki morfologi yang cukup kompleks. Di tengah-tengah tantangan ini, pendekatan berbasis *deep learning*, terutama yang menggunakan arsitektur *Convolutional Neural Networks* (CNNs), telah berkembang pesat dan menunjukkan potensi yang sangat besar dalam meningkatkan performa NER, terutama dalam menangani permasalahan representasi kata dan konteks entitas dalam teks yang beragam. CNNs bekerja dengan mempelajari pola-pola fitur dari data teks dan memiliki keunggulan dalam menangani representasi spasial, yang menjadikannya sangat cocok untuk tugas-tugas pengenalan entitas pada teks bahasa alami.

Pendekatan CNNs dalam NER untuk bahasa Indonesia semakin relevan untuk dipertimbangkan, mengingat penelitian-penelitian sebelumnya menunjukkan bahwa model-model berbasis CNN, khususnya dalam kombinasi dengan arsitektur *Long Short-Term Memory* (LSTM) *bidirectional*, mampu mencapai tingkat akurasi yang sangat memadai dalam mengidentifikasi dan mengklasifikasikan entitas dari teks bahasa Indonesia. Misalnya, penelitian yang dilakukan oleh [3] menunjukkan bahwa pendekatan *hibrida* antara LSTM dan CNN berhasil menunjukkan performa yang baik dalam mengekstrak informasi entitas seperti orang, organisasi, lokasi, dan acara dari artikel bahasa Indonesia. Dalam penelitian ini, model NER berhasil mengkategorikan entitas-entitas tersebut dengan tingkat akurasi yang cukup tinggi, yang membuktikan bahwa pendekatan ini dapat berfungsi secara efektif untuk bahasa Indonesia. Hal ini menjadi bukti bahwa arsitektur CNN yang digabungkan dengan LSTM mampu mengatasi tantangan unik yang dihadapi dalam pengembangan NER untuk bahasa Indonesia [4].

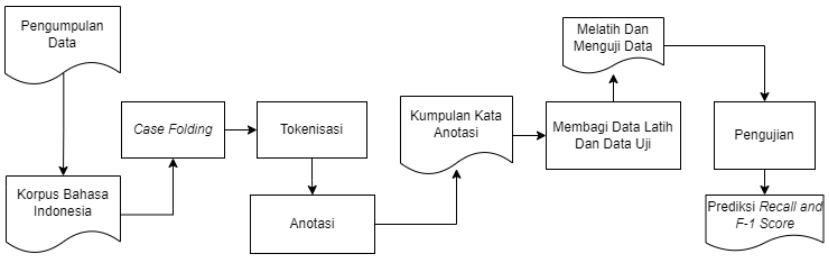
Lebih lanjut, penelitian oleh [5] mengungkapkan bahwa *embedding* kata, seperti *word2vec*, memainkan peran penting dalam mempermudah proses pelatihan model NER. *Embedding* kata memungkinkan model untuk memahami konteks semantik dari setiap kata dalam kalimat, sehingga membantu model dalam membedakan antara berbagai jenis entitas. Dalam penelitian tersebut, penggunaan arsitektur *Bidirectional* LSTM yang digabungkan dengan CNN dan *embedding word2vec* berhasil menghasilkan peningkatan performa yang signifikan. Dengan nilai F1 *score* sebesar 71,37%, model ini menunjukkan bahwa pendekatan ini mampu mengatasi variasi format teks dan kekayaan morfologis bahasa Indonesia, yang sering kali menjadi tantangan dalam tugas pengenalan entitas.

Selain itu, CNNs juga terbukti sangat efisien dalam menangani dataset teks yang tidak terstruktur. Sebagai contoh, penelitian oleh [6] menyoroti penerapan CNNs dalam pengenalan entitas pada teks tidak terstruktur, seperti teks yang dihasilkan oleh sistem konversi suara ke teks. Dataset yang tidak terstruktur sering kali menimbulkan kesulitan tambahan karena tata bahasa yang lebih longgar dan kesalahan transkripsi, namun CNNs menunjukkan kemampuan yang solid dalam memproses teks semacam itu. Dalam penelitian ini, menggabungkan beberapa modifikasi dataset serta algoritma *deep learning* seperti LSTM, CNN, dan GRU. Melalui beberapa skenario eksperimen, penelitian ini berhasil mencapai skor F1 hingga 71,04%, menunjukkan bahwa CNNs efektif dalam menangani variasi data dan format teks yang tidak terstruktur.

Secara keseluruhan, pendekatan berbasis CNNs dalam pengembangan NER untuk bahasa Indonesia menunjukkan hasil yang sangat menjanjikan. Berbagai penelitian yang telah dilakukan dalam beberapa tahun terakhir memperlihatkan bahwa performa NER dapat terus ditingkatkan melalui pengembangan teknik *embedding*, modifikasi dataset, dan penggunaan arsitektur deep learning yang lebih canggih. Dalam penelitian ini, kami akan mengeksplorasi lebih lanjut potensi CNNs untuk meningkatkan performa NER pada teks berbahasa Indonesia, khususnya dalam menghadapi tantangan-tantangan yang ada seperti keterbatasan data dan variasi format teks yang sering dijumpai pada bahasa alami.

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model Named Entity Recognition (NER) pada teks berbahasa Indonesia menggunakan arsitektur Convolutional Neural Networks (CNNs). Tujuan utama dari penelitian ini adalah untuk meningkatkan akurasi pengenalan entitas dalam berbagai kategori seperti nama orang, organisasi, lokasi, dan entitas lainnya, dengan fokus pada peningkatan performa model dalam menghadapi tantangan unik bahasa Indonesia, seperti kompleksitas morfologi dan keterbatasan dataset. Penelitian ini juga bertujuan untuk mengeksplorasi bagaimana teknik embedding kata, serta pendekatan berbasis deep learning lainnya, dapat digunakan untuk mengatasi variasi format dan konteks teks tidak terstruktur dalam bahasa Indonesia, serta memaksimalkan performa NER secara keseluruhan. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem NER yang lebih efisien dan akurat untuk teks berbahasa Indonesia.

2. METODE PENELITIAN



Gambar 1. Alur Penelitian

2.1 Pengumpulan Dataset

Pada tahap pertama, dilakukan pengumpulan dataset yang terdiri dari berbagai sumber teks berbahasa Indonesia yang beragam. Sumber-sumber yang digunakan meliputi artikel berita, dokumen formal, serta teks dari transkripsi suara ke teks. Pengambilan data dengan menggunakan *web scraping. Web scraping* adalah teknik yang digunakan untuk mendapatkan informasi dari situs web, dan *web scraping* akan mengekstrak data secara otomatis tanpa menyalinnya secara otomatis tanpa menyalinnya secara manual [7]. Pengumpulan data ini mencakup entitas yang dikelompokkan ke dalam beberapa kategori, seperti nama orang, organisasi, lokasi, tanggal, dan entitas lainnya. Tujuan pengumpulan dataset adalah untuk menciptakan data yang kaya dan representatif dari berbagai variasi bahasa Indonesia. Variasi ini mencakup teks terstruktur, yang biasanya lebih formal (misalnya artikel berita), serta teks tidak terstruktur, yang lebih informal (misalnya teks dari media sosial atau transkripsi percakapan). Adanya teks tidak terstruktur ini penting untuk memastikan bahwa model NER dapat bekerja pada kondisi bahasa Indonesia yang beragam, seperti yang ditemukan dalam teks percakapan sehari-hari atau tulisan informal lainnya.

2.2 Korpus Bahasa Indonesia

Korpus bahasa Indonesia adalah kumpulan teks atau ujaran dalam bahasa Indonesia yang dikumpulkan secara sistematis untuk tujuan penelitian linguistik. Korpus ini bisa berupa teks tertulis seperti buku, artikel, dan blog, atau rekaman lisan seperti wawancara dan siaran radio [8]. Dalam linguistik, korpus digunakan untuk menganalisis penggunaan bahasa secara nyata, membantu dalam penyusunan kamus, dan mendukung penelitian tentang struktur dan penggunaan bahasa. Di Indonesia, beberapa korpus yang terkenal adalah Korpus Indonesia (KOIN) dan Korpus Nusantara.

2.3 *Preprocessing*

Proses *preprocessing* yang dilakukan dalam penelitian ini adalah sebagai berikut:

a. *Case Folding*

   Case Folding dilakukan dengan tujuan menghilangkan tanda baca dan karakter yang tidak perlu pada *dataset* berbahasa Indonesia selain tanda titik (“.”), koma (“,”), garis miring (“/”), kurung (“()”), huruf “a” sampai “z”, dan karakter angka “0...9” serta menghilangkan *whitespace* atau karakter kosong pada data teks[4]

b. *Tokenization*

   Tokenization dilakukan untuk memecah teks menjadi token-token atau kata-kata terpisah. Ini adalah langkah pertama yang penting untuk memisahkan teks berdasarkan spasi atau tanda baca [9].

2.4 Pelebelan Data

Format *Stanford* BIO (*Begin, Inside, Outside*) digunakan dalam NER untuk melabeli token dalam sebuah teks berdasarkan apakah token tersebut merupakan awal, bagian lanjutan, atau di luar dari suatu entitas. Dalam format ini, *B-ENTITY* (*Begin*) menandai token yang memulai sebuah entitas, seperti *B-PERSON* untuk token pertama dari nama orang. *I-ENTITY* (*Inside*) digunakan untuk menandai token yang merupakan bagian lanjutan dari entitas yang sama, seperti *I-PERSON* untuk nama belakang orang yang terdiri dari beberapa kata. Sementara itu, O (*Outside*) menandai token yang tidak termasuk dalam entitas yang diakui, seperti kata-kata umum yang tidak terkait dengan entitas tertentu [10].

2.5 Pembagian Data

Hasil dari data yang telah di lebel kemudian data yang telah dilakukan pelebelan data akan di *spilt* menjadi *train data*, *test data*, *validation data* [11]. Dalam penelitian ini, kami telah melatih model dengan implementasi medan acak bersyarat yang disediakan oleh *Tensorflow*. Pembagian data dalam bentuk sampel menjadi data training dan data testing pada penelitian ini menggunakan library scikit-learn. Pembagian bagian dataset menjadi dua bagian, training dan testing menggunakan metode train-test split pada library python. Skenario ini bertujuan untuk menguji sebuah model pada kondisi dataset yang berbeda. Pada penelitian ini, kami menggunakan tiga skenario, yaitu 20% data testing dan 80% data training, 30% data testing dan 70% data training, dan 40% data testing dan 60% data pelatihan [12].

2.6 Prediksi

Dalam penelitian ini, kami mengukur nilai ketepatan, *recall*, dan F1 *score* untuk mengevaluasi kinerja model Indonesia-NER [13]. Metrik ini biasanya digunakan untuk mengevaluasi kinerja sistem klasifikasi *seperti Named Entity Recognition* (NER). Metrik ini mengukur seberapa efektif dan akurat sistem dalam mengidentifikasi dan mengklasifikasikan entitas dalam teks. Persentase prediksi positif yang benar yang dibuat oleh sistem klasifikasi sistem klasifikasi dibandingkan dengan jumlah prediksi positif total disebut presis. Presisi yang tinggi menunjukkan bahwa sistem tersebut baik dalam menemukan contoh prediksi positif yang benar dan tidak membuat prediksi positif yang salah. Persentase contoh positif asli yang diidentifikasi oleh sistem klasifikasi dibandingkan dengan total jumlah contoh positif dalam *dataset*. Persentase penarikan kembali yang tinggi menunjukkan bahwa sistem dapat menemukan sebagian besar atau semua contoh positif.

3. HASIL DAN PEMBAHASAN

Data penelitian ini diperoleh dari GitHub melalui repositori [ner-dataset-modified-dee] (https://github.com/ialfina/ner-dataset-modified-dee.git). Dataset ini berisi data yang dimodifikasi untuk Pengenalan Entitas Bernama (*Named Entity Recognition*) pada teks bahasa Indonesia. Data tersebut disesuaikan untuk mendukung berbagai eksperimen dan pengujian model NER. Penulis menggunakan dataset ini sebagai sumber utama dalam penelitian, di mana data tersebut telah melalui tahap pra-pemrosesan seperti tokenisasi dan pelabelan entitas, yang selanjutnya digunakan dalam proses pelatihan dan evaluasi model *machine learning*.

3.1 *Tokenization*

Hasil dari langkah tokenisasi dalam penelitian ini adalah kumpulan kata yang berhubungan dengan pemberitaan, yang terdiri dari 20.000 kata. Terdapat berbagai pendekatan yang dapat digunakan untuk proses tokenisasi, dan pemilihan pendekatan ini bergantung pada persyaratan serta batasan spesifik dari sistem Pengenalan Entitas Bernama (NER). Beberapa metode tokenisasi standar yang digunakan di antaranya adalah tokenisasi tingkat kata, yang membagi teks menjadi kata-kata individual, serta tokenisasi tingkat karakter, yang membagi teks menjadi karakter-karakter individual. Metode tokenisasi yang dipilih dalam penelitian ini disesuaikan dengan kebutuhan analisis teks pemberitaan untuk mencapai hasil yang optimal dalam pengenalan entitas.

Tabel 1. Contoh *Tokenization*

Dokumen	***Tokenization***	
Hakim Pengadilan Negeri (PN) Nunukan, Kalimantan Utara, memulai aksi mogok sidang sebagai bentuk solidaritas terhadap rekan-rekan hakim di seluruh Indonesia yang menuntut peningkatan kesejahteraan.	Hakim Pengadilan Negeri (PN) Nunukan , mogok sidang sebagai bentuk .	solidaritas terhadap rekan-rekan hakim di seluruh Indonesia yang menuntut peningkatan kesejahteraan

3.2 Anotasi

Seperti yang telah disebutkan sebelumnya, dalam penelitian ini, kami menggunakan set data yang terdiri dari 20.000 kata yang dianotasi secara manual. Beberapa tokenisasi standar termasuk tokenisasi tingkat kata, yang membagi teks menjadi kata-kata individual, dan tokenisasi tingkat karakter, yang membagi teks menjadi karakter individual [14]. Dalam anotasi dokumen, pelabelan kata dilakukan dengan menggunakan format stanford BIO, yang merupakan singkatan dari *Beginning* (B), *Inside* (I) dan *Outside* (O). Dalam format anotasi umum untuk menandai suatu entitas, dalam format BIO di mana "B" adalah awal dari entitas, "I" adalah entitas perantara atau masih terkait dengan entitas, dan "O" adalah entitas, dan "O" adalah kata yang tidak termasuk entitas bernama [15]. Contoh anotasi manual ditunjukkan pada Tabel 2.

Tabel 2. Contoh Anotasi Teks

Entitas	**Anotasi**
Hakim	B-PERSON
Pengadilan	B-ORGANIZATION
Negeri	I-ORGANIZATION
(	O
PN	B-ORGANIZATION
)	O

Nunukan	B-PLACE
,	O
Kalimantan	B-PLACE
Utara	I-PLACE

Pada Tabel 2, kami telah mengidentifikasi empat entitas dalam kalimat tersebut: Hakim, Pengadilan Negeri, Nunukan, dan Kalimantan Utara. "Hakim" adalah kata yang ditandai dengan label B-PERSON (lihat Tabel 2). "Pengadilan Negeri" adalah frasa dengan dua kata yang ditandai dengan B-ORGANIZATION dan I-ORGANIZATION. Nunukan dan Kalimantan Utara merupakan frasa yang menandakan tempat (place), dengan label B-PLACE dan I-PLACE untuk Kalimantan Utara. Dengan pelabelan ini, kita dapat melihat apakah sebuah kata berdiri sendiri atau merupakan bagian dari frasa yang lebih besar.

Dengan adanya format BIO, kita dapat mengurangi kebingungan dalam pengenalan kata atau frasa yang serupa namun memiliki makna berbeda. Sebagai contoh, kalimat "Hakim PN Nunukan" dan "Kalimantan Utara" sama-sama mengandung entitas yang berhubungan dengan tempat dan organisasi, namun dengan pelabelan yang jelas, perbedaan entitas tersebut dapat diidentifikasi dengan baik.

Pada penelitian ini, kami menggunakan *Convolutional Neural Networks* (CNNs) untuk menangani proses pengenalan entitas. CNNs mampu menangkap pola dari urutan kata dan frasa dalam kalimat melalui proses konvolusi, yang dapat mengekstrak fitur spasial dari teks. Jaringan CNNs bekerja dengan mendeteksi fitur penting dari setiap kata atau frasa berdasarkan konteks lokalnya, yang kemudian digunakan untuk memperbaiki klasifikasi entitas. CNNs tidak hanya fokus pada satu token (kata) secara individual, namun juga memperhatikan hubungan dengan token-token lainnya di sekitarnya.

Keunggulan CNNs dalam menangani pengenalan entitas adalah kemampuannya untuk menangkap pola-pola yang kompleks dan berulang di dalam teks, yang sangat berguna dalam mengenali perbedaan antara frasa atau kata yang mirip namun memiliki makna atau fungsi yang berbeda dalam konteks kalimat.

3.3 *Training*

Untuk melatih model, kami menggunakan CNNs yang disediakan oleh *tensorflow*. Kami memulai contoh model dan sesuaikan data pelatihan dengan metode fit. Skrip untuk melatih model Indonesia-NER dapat ditemukan di bawah ini.

```
history = model.fit(X, np.array(y), batch_size=32, epochs=5, validation_split=0.1)
```

Gambar 2. Melatih Model

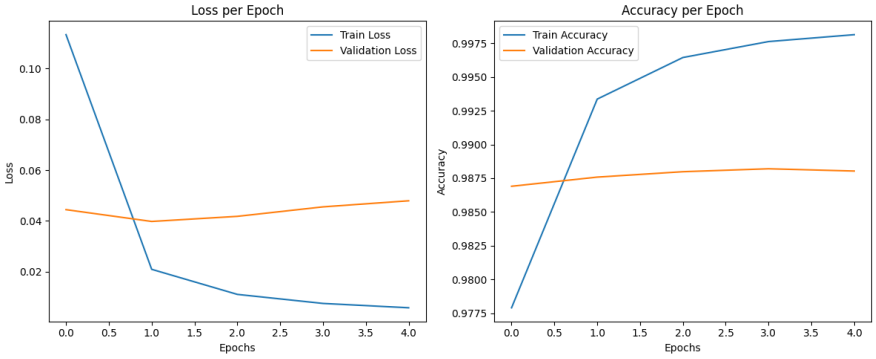
Menggunakan metode *fit()* untuk melatih model dengan data yang telah diproses. Pada proses ini, model akan membagi data menjadi set pelatihan dan set validasi secara acak berdasarkan parameter validation_split=0.1, yang berarti 10% dari data akan digunakan untuk validasi. Karena pembagian data ini dilakukan secara acak pada setiap *running*, maka jika proses pelatihan diulang, hasilnya bisa sedikit berbeda setiap kali. Selain itu, parameter seperti batch_size=32 dan *epochs*=5 mengatur bagaimana model akan belajar dari data. Dimana *batch_size=32* berarti model akan memproses 32 sampel sekaligus sebelum memperbarui bobotnya. *epochs*=5 menunjukkan bahwa model akan melewati seluruh dataset sebanyak 5 kali.

Penting untuk diingat bahwa model ini tidak menggunakan teknik seperti *cross-validation* yang membagi data secara konsisten dalam berbagai iterasi. Oleh karena itu, hasil evaluasi dapat berubah setiap kali pelatihan dilakukan karena pembagian set pelatihan dan validasi yang acak.

Namun, struktur dan performa model CNN yang digunakan tidak akan berubah, dan jika Anda ingin mendapatkan hasil yang lebih stabil atau menghindari variasi yang dihasilkan oleh pembagian acak ini, Anda bisa menggunakan metode pembagian data yang tetap (misalnya dengan *train_test_split()* yang disimpan dengan seed tetap).

3.4 *Testing*

Pada penelitian ini, pengujian performa model Indonesia-NER dilakukan dengan lima percobaan atau skenario dengan menggunakan split train/test.



Gambar 3 Grafik Pelatihan Model

Grafik yang ditampilkan menunjukkan hasil dari pelatihan model dalam bentuk *Loss* per *Epoch* dan *Accuracy* per *Epoch*. Pada grafik Loss, *Train Loss* (garis biru) menunjukkan bahwa nilai kehilangan pada data pelatihan menurun drastis, terutama pada awal pelatihan, dan cenderung stabil di bawah 0.05 setelah epoch pertama. Ini menunjukkan bahwa model berhasil belajar dengan baik dari data pelatihan. Sebaliknya, *Validation Loss* (garis oranye) tetap relatif stabil tanpa penurunan yang signifikan, yang bisa mengindikasikan bahwa model mungkin tidak mampu belajar dengan baik dari data validasi atau terdapat perbedaan antara data pelatihan dan validasi.
Pada grafik *Accuracy*, *Train Accuracy* menunjukkan peningkatan yang tajam dan hampir mencapai 100% pada epoch akhir, menunjukkan kinerja model yang sangat baik pada data pelatihan. Namun, *Validation Accuracy* cenderung lebih rendah dan menunjukkan peningkatan yang lambat, yang bisa menjadi tanda adanya *overfitting*; model belajar terlalu baik pada data pelatihan namun tidak dapat menggeneralisasi dengan baik pada data baru. Dengan demikian, meskipun model menunjukkan performa yang tinggi pada data pelatihan, perbedaan signifikan antara akurasi pelatihan dan validasi menyarankan perlunya langkah-langkah untuk meningkatkan generalisasi, seperti menambah data, menggunakan teknik regulasi, atau memodifikasi arsitektur model.

4. DISKUSI

Berdasarkan hasil pengujian model yang telah dilatih, berikut adalah nilai **precision**, **recall**, dan **f1-score** untuk setiap kategori:

Tabel 3. Hasil Evaluasi Model

Label	Precision	Recall	F1-Score	Support
Place	0.96	0.96	0.96	22,229
Organisation	0.92	0.82	0.88	4,766
Person	0.97	0.94	0.95	19,733
O	0.94	0.97	1.00	550,766
Total	1.00	1.00	1.00	1,426,506

Berdasarkan hasil pengujian model yang diberikan, berikut adalah nilai *precision*, *recall*, dan f1-*score* untuk setiap kategori. Untuk kategori *Place*, model mencapai *precision* dan *recall* masing-masing sebesar 0.96, menghasilkan *f1-score* 0.96 dengan support 22,229. Kategori *Organisation* memiliki *precision* 0.92 dan *recall* 0.85, yang memberikan *f1-score* 0.88 dengan *support* 4,766. Sementara itu, kategori *Person* menunjukkan hasil yang sangat baik dengan *precision* 0.97, *recall* 0.94, dan *f1-score* 0.95, didukung oleh 19,733 entitas. Label O (yang menunjukkan entitas lain) memperoleh nilai sempurna dengan *precision*, *recall*, dan *f1-score* masing-masing 1.00, dengan *support* mencapai 550,766. Secara total, model menunjukkan akurasi 1.00 (100%), *dengan macro average precision* sebesar 0.97, *recall* 0.95, dan *f1-score* 0.96. Selain itu, *weighted average* juga mencatatkan nilai optimal, dengan *precision*, *recall*, dan *f1-score* masing-masing 1.00. Hasil ini menunjukkan bahwa model memiliki performa yang sangat baik, terutama dengan nilai *F1-score* yang tinggi, terutama untuk kategori umum (label O). Kategori Person, Place, dan Organisation juga menunjukkan performa yang baik, meskipun terdapat sedikit perbedaan dalam *recall* untuk kategori *Organisation*.

5. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa pengembangan model Named Entity Recognition (NER) berbasis arsitektur *Convolutional Neural Networks* (CNNs) untuk teks berbahasa Indonesia menunjukkan hasil yang sangat menjanjikan. Model ini berhasil mencapai akurasi total sebesar 100% dan menghasilkan nilai *precision, recall, dan F1-score* yang tinggi untuk berbagai kategori entitas, termasuk kategori *Place, Organisation*, dan *Person*. Meskipun ada perbedaan kecil dalam nilai *recall* untuk kategori *Organisation*, secara keseluruhan,

model mampu mengenali dan mengklasifikasikan entitas dengan baik, menunjukkan kemampuannya dalam menangani tantangan unik yang dihadapi dalam pemrosesan bahasa Indonesia. Hasil evaluasi model menunjukkan bahwa pendekatan berbasis CNN, terutama ketika dipadukan dengan teknik *embedding* dan modifikasi dataset, dapat meningkatkan performa NER dan membantu mengatasi kompleksitas morfologi serta variasi format teks yang sering ditemui. Dengan demikian, penelitian ini memberikan kontribusi signifikan terhadap pengembangan sistem NER yang lebih efisien dan akurat untuk bahasa Indonesia, serta membuka jalan bagi penelitian lebih lanjut di bidang ini.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Ibu Rini Meiyanti selaku pembimbing dalam penyusunan penelitian ini. Bimbingan, saran, dan dukungan Ibu sangat berarti dalam proses penyelesaian penelitian ini. Selain itu, penulis juga mengucapkan terima kasih kepada penyedia dataset yang telah memberikan akses kepada data yang diperlukan untuk penelitian ini, sehingga memungkinkan penulis untuk melakukan eksperimen dan analisis yang mendalam. Tanpa dukungan dan kontribusi dari kedua pihak tersebut, penelitian ini tidak akan dapat terlaksana dengan baik.

DAFTAR PUSTAKA

[1] N. Yusliani, M. R. P. Sufa, A. Firdaus, Abdiansah, and S. Yoppy, "Named-Entity Recognition Pada Teks Berbahasa Indonesia Menggunakan Metode Hidden Markov Model Dan Part-of-Speech Tagging," *Linguist. Komputasional*, vol. 4, no. 1, pp. 13–20, 2020.

[2] D. S. Rachmad, "Review Named Entity Recognition dengan Menggunakan Machine Learning," *J. Sains dan Inf.*, vol. 6, no. 1, pp. 28–33, 2020, doi: https://doi.org/10.34128/jsi.v6i1.204.

[3] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko, "Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs," *Procedia Comput. Sci.*, vol. 135, pp. 425–432, 2018, doi: https://doi.org/10.1016/j.procs.2018.08.193.

[4] A. A. Putra and R. Kurniawan, "Bidirectional LSTM-Cnns Untuk Ekstraksi Entity Lokasi Kebakaran Pada Berita Online Berbahasa Indonesia," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, pp. 319–327, 2021, doi: 10.34123/semnasoffstat.v2020i1.601.

[5] S. Sukardi, M. Susanty, A. Irawan, and Randi Fermana Putra, "Low Complexity Named-Entity Recognition for Indonesian Language using BiLSTM-CNNs," *IEEE*, 2021, doi: https://doi.org/10.1109/ICOIACT50329.2020.9331989.

[6] N. S. Azzahra, M. O. Ibrohim, J. Fahmi, B. F. Apriyanto, and O. Riandi, "Developing Name Entity Recognition for Structured and Unstructured Text Formatting Dataset," *IEEE*, 2020, doi: DOI: 10.1109/ICIC50835.2020.9288566.

[7] Q. Liu, R. Yahyapour, H. Liu, and Y. Hu, "A novel combining method of dynamic and static web crawler with parallel computing," *Multimed Tools Appl*, vol. 83, 2024, doi: https://doi.org/10.1007/s11042-023-17925-y.

[8] J. Bata, "#AkuGalau: Korpus Bahasa Indonesia untuk Deteksi Emosi dari Teks," *J. Elektro*, vol. 12, no. 2, pp. 103–110, 2019, [Online]. Available: https://mx2.atmajaya.ac.id/index.php/JTE/article/view/1218

[9] N. F. Widiyanti, H. T. Sukmana, K. Hulliyah, D. Khairani, and L. K. Oh, "Improving Indonesian Named Entity Recognition for Domain Zakat Using Conditional Random Fields," *J. Online Inform.*, vol. 8, no. 2, pp. 131–138, 2023, doi: 10.15575/join.v8i2.898.

[10] R. Rifani, M. A. Bijaksana, and Asror Ibnu, "Named Entity Recognition for an Indonesian Based Language Tweet using Multinomial Naive Bayes Classifier," *Ind. J. Comput.*, vol. 4, no. 2, pp. 119–126, 2019, doi: 10.21108/indojc.2019.4.2.330.

[11] D. Christianto, E. Siswanto, and R. Chaniago, "Penggunaan Named Entity Recognition dan Artificial Intelligence Markup Language untuk Penerapan Chatbot Berbasis Teks," *J. Telemat.*, vol. 10, no. 2, pp. 61–68, 2019, doi: 10.61769/telematika.v10i2.130.

[12] L. FAHADRA, "AYANAN INFORMASI TUGAS AKHIR MENGGUNAKAN METODE LONG SHORT-TERM MEMORY (LSTM)," 2023. [Online]. Available: http://repository.unissula.ac.id/id/eprint/32030%0A

[13] B. S. JATI, Widyawan, and M. N. Rizal, "Model Named Entity Recognition Multilingual Untuk Sistem Tanya Jawab Asuransi Kesehatan Nasional," 2020.

[14] A. Sinaga and S. N. Pandapotan, "Analisis Perbandingan Akurasi Dan Waktu Proses Algoritma Stemming Arifin-Setiono Dan Nazief-Adriani Pada Dokumen Teks Bahasa Indonesia," *Univ. Telkom*, 2023, doi: https://doi.org/10.46984/sebatik.v27i1.2072.

[15] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named Entity Recognition and Relation Detection for Biomedical Information Extraction," *Front. Cell Dev. Biol.*, vol. 8, Aug. 2020, doi: 10.3389/fcell.2020.00673.