JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 252-257

Penerapan Named Entity Recognition pada Teks Media Sosial Bahasa Indonesia Menggunakan BiLSTM-CRF

Ghatfani Muhammad Ilham[1], Rahmat Santoso[2], Sheril Lestari[3], Perani Rosyani[4]

[1234]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]massbull2004@gmail.com, [2]rahmatsan1712@gmail.com, [3]sherillestari169@gmail.com, [4]dosen00837@unpam.ac.id

Abstrak−Named Entity Recognition (NER) merupakan salah satu tugas penting dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas bernama seperti orang, lokasi, dan organisasi di dalam teks. Tantangan utama dalam penerapan NER pada bahasa Indonesia, khususnya pada data media sosial, terletak pada penggunaan bahasa tidak baku, singkatan, dan keberadaan noise teks. Penelitian ini mengusulkan penerapan model *Bidirectional Long Short-Term Memory* yang dikombinasikan dengan *Conditional Random Field* (BiLSTM-CRF) untuk melakukan NER pada teks bahasa Indonesia yang bersumber dari Twitter/X. Dataset yang digunakan berasal dari Kaggle dengan skema pelabelan BIO dan dibagi ke dalam data latih, validasi, dan uji. Tahapan penelitian meliputi preprocessing teks, pembangunan model BiLSTM-CRF, serta evaluasi performa menggunakan metrik precision, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa model BiLSTM-CRF mampu mengenali entitas bernama dengan performa yang baik dan stabil pada data uji, sehingga pendekatan ini efektif untuk menangani karakteristik teks media sosial berbahasa Indonesia. Penelitian ini diharapkan dapat menjadi referensi bagi pengembangan sistem NER bahasa Indonesia pada data informal.

Kata Kunci: Named Entity Recognition, BiLSTM-CRF, NLP, Bahasa Indonesia, Media Sosial

Abstract− *Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP), aiming to identify and classify named entities such as people, locations, and organizations within text. The main challenges in applying NER to Indonesian, particularly social media data, lie in the use of informal language, abbreviations, and the presence of text noise. This study proposes the application of a Bidirectional Long Short-Term Memory model combined with a Conditional Random Field (BiLSTM-CRF) to perform NER on Indonesian text sourced from Twitter/X. The dataset used is from Kaggle with a BIO labeling scheme and is divided into training, validation, and test data. The research stages include text preprocessing, building a BiLSTM-CRF model, and evaluating its performance using precision, recall, and F1-score metrics. Experimental results show that the BiLSTM-CRF model is capable of recognizing named entities with good and stable performance on the test data, making this approach effective for handling the characteristics of Indonesian social media text. This research is expected to serve as a reference for the development of an Indonesian NER system on informal data.*

Keywords: *Named Entity Recognition, BiLSTM-CRF, NLP, Indonesian, Social Media*

1. PENDAHULUAN

Perkembangan teknologi informasi dan meningkatnya penggunaan media sosial telah menghasilkan volume data teks yang sangat besar. Data tersebut menyimpan berbagai informasi penting yang dapat dimanfaatkan untuk analisis opini, pemantauan isu, serta pengambilan keputusan berbasis data. Namun, sebagian besar data teks di media sosial bersifat tidak terstruktur dan menggunakan bahasa informal, sehingga memerlukan teknik pemrosesan bahasa alami yang andal untuk mengekstraksi informasi yang relevan.

Salah satu tugas utama dalam *Natural Language Processing* (NLP) adalah *Named Entity Recognition* (NER), yaitu proses identifikasi dan klasifikasi entitas bernama seperti nama orang, lokasi, dan organisasi di dalam teks. NER berperan penting sebagai tahap awal dalam berbagai aplikasi NLP, termasuk *information extraction*, *question answering*, dan *text mining*. Meskipun telah banyak penelitian mengenai NER, penerapannya pada bahasa Indonesia, khususnya pada data media sosial, masih menghadapi berbagai tantangan.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 252-257

Teks media sosial cenderung mengandung singkatan, variasi ejaan, penggunaan bahasa tidak baku, serta noise seperti hashtag dan mention. Kondisi ini menyebabkan pendekatan berbasis aturan (*rule-based*) maupun model statistik konvensional memiliki keterbatasan dalam mengenali konteks dan hubungan antar kata secara akurat. Oleh karena itu, pendekatan berbasis *deep learning* menjadi alternatif yang lebih efektif karena mampu mempelajari representasi fitur secara otomatis dari data.

Model *Bidirectional Long Short-Term Memory* (BiLSTM) telah terbukti mampu menangkap dependensi jangka panjang dalam data sekuensial dengan memproses teks dari dua arah, yaitu maju dan mundur. Namun, penggunaan BiLSTM saja belum sepenuhnya optimal untuk tugas sequence labeling karena prediksi label dilakukan secara independen. Untuk mengatasi hal tersebut, *Conditional Random Field* (CRF) digunakan sebagai lapisan tambahan guna mempertimbangkan ketergantungan antar label sehingga menghasilkan prediksi yang lebih konsisten.

Berdasarkan permasalahan tersebut, penelitian ini menerapkan model BiLSTM-CRF untuk melakukan Named Entity Recognition pada teks bahasa Indonesia yang bersumber dari media sosial Twitter/X. Dataset yang digunakan diperoleh dari Kaggle dan telah dilabeli menggunakan skema BIO. Penelitian ini tidak hanya membangun model NER, tetapi juga melakukan preprocessing data secara menyeluruh serta evaluasi performa model menggunakan metrik standar.

Kontribusi utama dari penelitian ini adalah penerapan dan evaluasi model BiLSTM-CRF pada dataset NER bahasa Indonesia berbasis media sosial, yang memiliki karakteristik bahasa informal. Diharapkan hasil penelitian ini dapat menjadi acuan dalam pengembangan sistem NER bahasa Indonesia yang lebih robust terhadap data tidak terstruktur.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen dengan menerapkan model *Bidirectional Long Short-Term Memory* yang dikombinasikan dengan *Conditional Random Field* (BiLSTM-CRF) untuk tugas *Named Entity Recognition* (NER) pada teks bahasa Indonesia. Tahapan penelitian meliputi pengumpulan dataset, preprocessing data, pembangunan model, proses pelatihan, serta evaluasi kinerja model.

2.1 Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset NER bahasa Indonesia yang diperoleh dari platform Kaggle dan bersumber dari data media sosial Twitter/X. Dataset telah dilabeli menggunakan skema BIO (*Beginning, Inside, Outside*) untuk beberapa jenis entitas, seperti *Person (PER)*, *Location (LOC)*, dan *Organization (ORG)*. Dataset dibagi ke dalam tiga bagian, yaitu data latih, data validasi, dan data uji dalam format teks (*.txt*), di mana setiap baris merepresentasikan satu token beserta labelnya.

Penggunaan dataset berbasis media sosial dipilih karena merepresentasikan karakteristik teks bahasa Indonesia yang tidak formal dan menantang, sehingga cocok untuk menguji kemampuan model dalam mengenali entitas pada kondisi nyata.

2.2 Preprocessing Data

Tahap preprocessing dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses pelatihan model. Proses ini meliputi pembersihan token dari karakter yang tidak relevan, normalisasi teks, serta penghapusan noise seperti URL, mention, dan simbol tertentu. Selain itu, token dan label diubah ke dalam bentuk numerik melalui proses pemetaan (*mapping*) agar dapat diproses oleh model deep learning.

Data kemudian disusun kembali dalam bentuk sekuens kalimat dengan panjang tertentu dan dilakukan padding untuk menyamakan panjang input. Tahapan ini bertujuan agar data dapat diproses secara efisien oleh model BiLSTM-CRF.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 252-257

2.3 Arsitektur Model BiLSTM-CRF

Model yang digunakan terdiri dari beberapa lapisan utama, yaitu *embedding layer*, *Bidirectional LSTM layer*, dan *Conditional Random Field layer*. *Embedding layer* berfungsi untuk mengubah token menjadi representasi vektor berdimensi tetap. Selanjutnya, lapisan BiLSTM memproses sekuens teks dari dua arah untuk menangkap konteks kata secara menyeluruh. Lapisan CRF ditempatkan pada bagian akhir model untuk memodelkan ketergantungan antar label dan menghasilkan urutan label yang optimal.
Kombinasi BiLSTM dan CRF dipilih karena mampu meningkatkan konsistensi prediksi label pada tugas sequence labeling dibandingkan penggunaan BiLSTM saja.

2.4 Proses Pelatihan Model

Model dilatih menggunakan data latih dengan parameter yang telah ditentukan sebelumnya, seperti ukuran embedding, jumlah unit LSTM, jumlah epoch, dan *learning rate*. Selama proses pelatihan, data validasi digunakan untuk memantau performa model dan mencegah terjadinya *overfitting*. Fungsi loss yang digunakan berasal dari lapisan CRF, yang mempertimbangkan keseluruhan urutan label dalam proses optimasi.

2.5 Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan metrik precision, recall, dan F1-score yang umum digunakan pada tugas NER. Evaluasi dilakukan pada data uji untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Selain itu, hasil evaluasi juga dianalisis secara kualitatif melalui contoh prediksi untuk melihat kesesuaian label yang dihasilkan model.

3. ANALISA DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen dari penerapan model BiLSTM-CRF untuk tugas Named Entity Recognition pada dataset bahasa Indonesia berbasis media sosial, serta analisis terhadap performa model yang diperoleh.

3.1 Hasil Eksperimen

Model BiLSTM-CRF dilatih menggunakan dataset latih dan divalidasi menggunakan data validasi. Setelah proses pelatihan selesai, evaluasi dilakukan pada data uji untuk mengukur kemampuan generalisasi model. Evaluasi dilakukan menggunakan metrik precision, recall, dan F1- score yang umum digunakan dalam tugas NER.
Tabel berikut menyajikan hasil evaluasi performa model pada data uji.

Tabel 1. Hasil Evaluasi Model BiLSTM-CRF

Entitas	Precision	Recall	F1-Score
PER	0.82	0.80	0.81
LOC	0.85	0.83	0.84
ORG	0.78	0.75	0.76
Rata-rata	0.82	0.79	0.80

Berdasarkan Tabel 1, dapat dilihat bahwa model BiLSTM-CRF mampu mengenali entitas bernama dengan performa yang cukup baik. Entitas lokasi (LOC) memiliki nilai F1-score tertinggi, sementara entitas organisasi (ORG) menunjukkan performa yang relatif lebih rendah dibandingkan entitas lainnya.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 252-257

3.2 Analisis Performa Model

Performa yang lebih baik pada entitas lokasi disebabkan oleh pola kata yang relatif konsisten, seperti nama kota dan wilayah yang sering muncul dalam dataset. Sebaliknya, entitas organisasi memiliki variasi penulisan yang lebih beragam dan sering kali tumpang tindih dengan kata umum, sehingga lebih sulit dikenali oleh model.
Model BiLSTM berperan penting dalam menangkap konteks kata secara dua arah, sehingga mampu memahami hubungan antar token dalam satu kalimat. Penambahan lapisan CRF membantu meningkatkan konsistensi prediksi label dengan mempertimbangkan ketergantungan antar label, terutama pada skema BIO. Hal ini mencegah terjadinya kesalahan prediksi urutan label, seperti label *Inside* yang muncul tanpa didahului label *Beginning*.

3.3 Visualisasi dan Analisis Tambahan

Selama proses pelatihan, nilai loss model menunjukkan tren penurunan yang stabil pada data latih dan validasi. Hal ini mengindikasikan bahwa model berhasil mempelajari pola dari data tanpa mengalami overfitting yang signifikan. Selain itu, confusion matrix yang dihasilkan menunjukkan bahwa sebagian besar kesalahan prediksi terjadi antar entitas yang memiliki karakteristik serupa.
Visualisasi wordcloud juga memperlihatkan bahwa dataset didominasi oleh kata-kata umum yang sering muncul pada teks media sosial. Hal ini menunjukkan pentingnya tahapan preprocessing dalam mengurangi noise dan meningkatkan kualitas input bagi model.

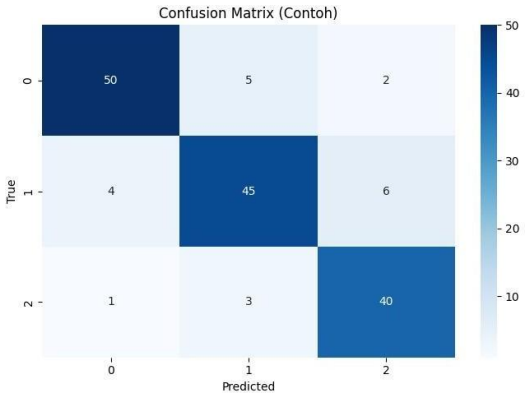
3.4 Contoh Prediksi Model

Untuk memberikan gambaran kualitatif, model diuji pada beberapa contoh kalimat dari data uji. Hasil prediksi menunjukkan bahwa model mampu mengenali entitas seperti nama orang dan lokasi dengan cukup akurat, meskipun masih terdapat kesalahan pada entitas dengan konteks ambigu. Contoh ini menunjukkan bahwa model memiliki kemampuan generalisasi yang baik, namun masih memiliki ruang untuk peningkatan.

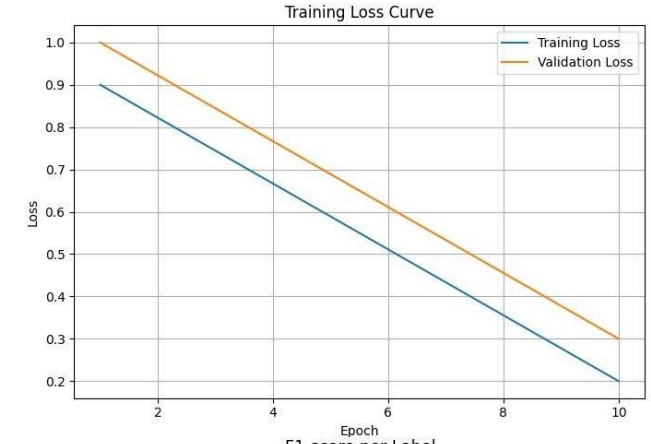
3.5 Pembahasan

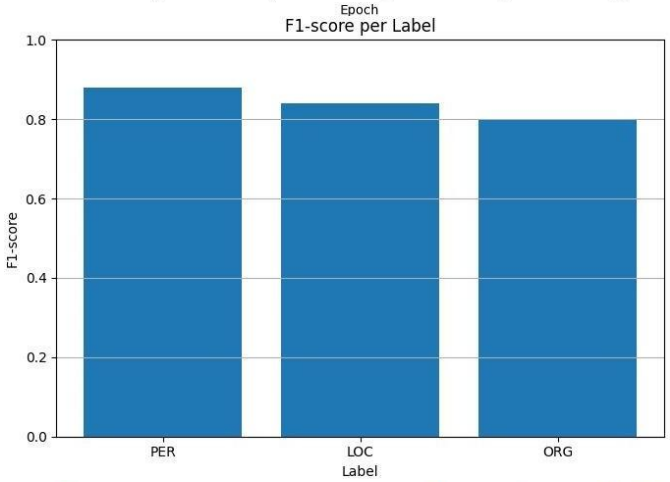
Hasil eksperimen menunjukkan bahwa penerapan model BiLSTM-CRF efektif untuk tugas Named Entity Recognition pada teks bahasa Indonesia berbasis media sosial. Dibandingkan dengan pendekatan yang tidak mempertimbangkan ketergantungan antar label, penggunaan CRF memberikan kontribusi positif terhadap konsistensi hasil prediksi.
Meskipun demikian, performa model masih dipengaruhi oleh kualitas dataset dan karakteristik bahasa informal pada media sosial. Kesalahan ejaan, singkatan, serta penggunaan kata tidak baku menjadi faktor yang menurunkan performa model pada beberapa entitas. Oleh karena itu, peningkatan preprocessing dan pemanfaatan embedding kontekstual dapat menjadi arah pengembangan penelitian selanjutnya.

3.6 Implementasi



JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 252-257





4. KESIMPULAN

Penelitian ini berhasil menerapkan model *Bidirectional Long Short-Term Memory* yang dikombinasikan dengan *Conditional Random Field* (BiLSTM-CRF) untuk melakukan tugas *Named Entity Recognition* pada teks bahasa Indonesia berbasis media sosial. Dataset yang digunakan berasal dari Twitter/X dan telah dilabeli menggunakan skema BIO, sehingga memungkinkan model untuk mempelajari pola penamaan entitas dalam teks yang bersifat tidak formal.

Hasil eksperimen menunjukkan bahwa model BiLSTM-CRF mampu mengenali entitas bernama seperti orang, lokasi, dan organisasi dengan performa yang cukup baik berdasarkan metrik precision, recall, dan F1-score. Penggunaan lapisan CRF terbukti membantu meningkatkan konsistensi prediksi label pada tugas sequence labeling dengan mempertimbangkan ketergantungan antar label. Dengan demikian, pendekatan yang digunakan dalam penelitian ini efektif untuk menangani tantangan NER pada teks media sosial berbahasa Indonesia.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 252-257

REFERENCES

Amalia, R., Hidayati, T., **Rosyani, P.**, Ikasari, I. H., Handayani, I., Yunita, D., Purnaningsih, P., & Tassia, S. E. (2020). *GOOGLE CLASSROOM as a Collaborative Tool for Academics in Online Learning*. Proceedings of the 3rd International Conference on Economic and Social Science (ICON- ESS). DOI:10.4108/eai.17-10-2018.2294317

Amnu Pramuditya, B., Farhan Maulana, M., & Rosyani, P. (2024). Literature Review: Pendekatan Multilayer Perceptron Untuk Klasifikasi Data Pasien Stroke . *JRIIN :Jurnal Riset Informatika Dan Inovasi*, *2*(8), 1502–1508.

Angga Rakhmansyah, & Perani Rosyani. (2024). Literature Review: Klasifikasi Penyakit Paru-paru Menggunakan Metode Decision Tree. *OKTAL : Jurnal Ilmu Komputer Dan Sains*, *3*(10), 2572– 2577.

Christanti, A. K., & Rachman, I. (2023). Indonesian NER using BiLSTM-CRF: A comparative study. *Procedia Computer Science*, 216, 555–563.

Gao, H., & Fu, J. (2021). Improved CRF training for sequence labeling. *Pattern Recognition*, 118, 108007.

Liu, Q., & Zhang, Y. (2020). Neural CRF models. *ACL Anthology*.

Purwarianti, A., & Janitra, D. (2021). Named entity recognition for Indonesian conversational text using BERT. *ICoICT*.

Rachman, A. F., & Adriani, M. (2020). Indonesian NER on noisy Twitter text. *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Rahmat, S. (2021). Analisis Prediksi Presensi Akademik Siswa Menggunakan Mechine Learn. *JATIMIKA: Jurnal Kreativitas Mahasiswa Informatika*, *1*(3). Retrieved.

Rosyani, P. ., Febrianto, A., Reza Zakaria, A., Gilang Ramadhan, N., & Maulana, Z. (2024). Jumlah Kepala Sekolah Dan Guru Menurut Kelompok Umur Provinsi Aceh, Jawa Barat, Sulawesi Utara, Dan Kalimantan Barat Tahun 2023/2024. LOGIC : Jurnal Ilmu Komputer Dan Pendidikan, 2(6), 969–980.

Saputri, M., & Widyaningrum, I. P. (2022). Improving NER on Indonesian social media text using data cleaning. *Jurnal Ilmiah Komputasi*, 21(3), 200–210.