ANALISIS PARAMETER TERBAIK DALAM DETEKSI KESALAHAN LABEL I-DIS PADA *NAMED ENTITY RECOGNITION* JUDUL BERITA PENYAKIT MENULAR BERBASIS *HYBRID* INDOBERT–BILSTM–CRF

Skripsi

Oleh

ANITA CAROLINE GUNAWAN
NPM. 2217031074

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2026

ABSTRACT

ANALYSIS OF OPTIMAL PARAMETERS IN DETECTING I-DIS LABEL ERRORS IN NAMED ENTITY RECOGNITION OF INFECTIOUS DISEASE NEWS HEADLINES BASED ON A HYBRID INDOBERT–BILSTM–CRF MODEL

By

Anita Caroline Gunawan

Entity extraction in health news texts is an important component in supporting the automated analysis of infectious disease information. However, the characteristics of news headlines, which are concise and context-dense, pose challenges for entity labeling using the Named Entity Recognition (NER) approach. This study aims to develop an NER model based on a hybrid IndoBERT–BiLSTM–CRF architecture by integrating Part-of-Speech (POS) tagging features using Stanza and FastText embeddings to recognize disease entities in Indonesian news headlines, determine the optimal hyperparameter configuration, and analyze labeling errors, particularly in the I-DIS label. The research method includes data preprocessing, annotation using the BIO scheme, model training with several hyperparameter combinations, and evaluation using accuracy, precision, recall, F1-score, confusion matrix, and ROC–AUC metrics. The results show that the best-performing model achieves an accuracy of 98.26% and a weighted F1-score of 98.28%, and is capable of recognizing disease entities with an accuracy of approximately 99%. The optimal configuration is obtained with a hidden dimension of 256, one BiLSTM layer, a dropout rate of 0.22, a learning rate of 2.53×10^{-5}, and a batch size of 16. Error analysis indicates that the primary weakness lies in the I-DIS label due to data imbalance; however, it does not significantly affect the overall model performance in detecting disease entities. Overall, the hybrid IndoBERT–BiLSTM–CRF model, supported by POS tagging (Stanza) and FastText embeddings, proves to be effective in recognizing disease entities in Indonesian news headlines and demonstrates competitive performance in NER tasks within the healthcare domain.

Keywords: Named Entity Recognition, IndoBERT–BiLSTM-CRF, POS Tagging, FastText Embedding, Error Analysis.

ABSTRAK

ANALISIS PARAMETER TERBAIK DALAM DETEKSI KESALAHAN LABEL I-DIS PADA *NAMED ENTITY RECOGNITION* JUDUL BERITA PENYAKIT MENULAR BERBASIS *HYBRID* INDOBERT–BILSTM–CRF

Oleh

Anita Caroline Gunawan

Ekstraksi entitas pada teks berita kesehatan merupakan komponen penting dalam mendukung analisis otomatis informasi penyakit menular. Namun, karakteristik judul berita yang singkat dan padat konteks menimbulkan tantangan dalam pelabelan entitas menggunakan pendekatan *Named Entity Recognition* (NER). Penelitian ini bertujuan membangun model NER berbasis arsitektur hibrida IndoBERT–BiLSTM–CRF dengan integrasi fitur POS *tagging* menggunakan Stanza dan *embedding* FastText untuk mengenali entitas penyakit pada judul berita berbahasa Indonesia, menentukan konfigurasi *hyperparameter* terbaik, serta menganalisis kesalahan pelabelan, khususnya pada label I-DIS. Metode penelitian meliputi praproses data, anotasi dengan skema BIO, pelatihan model menggunakan beberapa kombinasi *hyperparameter*, dan evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, F1-*score*, *confusion matrix*, dan ROC–AUC. Hasil penelitian menunjukkan bahwa model terbaik mencapai *accuracy* sebesar 98,26% dan *weighted* F1-*score* sebesar 98,28%, serta mampu mengenali entitas penyakit dengan tingkat ketepatan sekitar 99%. Konfigurasi optimal diperoleh pada *hidden dimension* 256, satu lapisan BiLSTM, *dropout* 0,22, *learning rate* $2{,}53 \times 10^{-5}$, dan *batch size* 16. Analisis kesalahan menunjukkan bahwa kelemahan utama terdapat pada label I-DIS akibat ketidakseimbangan distribusi data, namun tidak memberikan dampak signifikan terhadap kinerja model dalam mendeteksi entitas penyakit secara keseluruhan. Secara keseluruhan, model hibrida IndoBERT–BiLSTM–CRF dengan dukungan fitur POS *tagging* (Stanza) serta *embedding* FastText terbukti efektif dalam mengenali entitas penyakit pada judul berita berbahasa Indonesia dan menunjukkan kinerja yang kompetitif pada tugas NER di domain kesehatan.

Kata-kata kunci: *Named Entity Recognition*, IndoBERT–BiLSTM-CRF, POS *Tagging*, FastText *Embedding*, *Error Analysis*.

ANALISIS PARAMETER TERBAIK DALAM DETEKSI KESALAHAN LABEL I-DIS PADA *NAMED ENTITY RECOGNITION* JUDUL BERITA PENYAKIT MENULAR BERBASIS *HYBRID* INDOBERT–BILSTM–CRF

ANITA CAROLINE GUNAWAN

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2026

Judul Skripsi	:	**ANALISIS PARAMETER TERBAIK DALAM DETEKSI KESALAHAN LABEL I-DIS PADA *NAMED ENTITY RECOGNITION* JUDUL BERITA PENYAKIT MENULAR BERBASIS *HYBRID* INDOBERT–BILSTM–CRF**
Nama Mahasiswa	:	Anita Caroline Gunawan
Nomor Pokok Mahasiswa	:	**2217031074**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Dian Kurniasari, S.Si.,M.Sc.	**Dr. Purnomo Husnul Khotimah, M.T.**
NIP. 196903051996032001	NIP. 198003232005022002

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Dian Kurniasari, S.Si.,M.Sc.** .....

Sekretaris : **Dr. Purnomo Husnul Khotimah, M.T.** .....

Penguji

Bukan Pembimbing : **Prof. Dra. Wamiliana, MA.,Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.

NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **13 April 2026**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Anita Caroline Gunawan**
Nomor Pokok Mahasiswa	:	**2217031074**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Analisis Parameter Terbaik dalam Deteksi Kesalahan Label I-DIS pada *Named Entity Recognition* Judul Berita Penyakit Menular Berbasis *Hybrid* IndoBERT–BiLSTM–CRF**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 13 April 2026

Penulis

Anita Caroline Gunawan

RIWAYAT HIDUP

Penulis memiliki nama lengkap Anita Caroline Gunawan, lahir pada 3 Agustus 2004. Penulis merupakan anak pertama dari tiga bersaudara, dari pasangan Bapak Gunawan Malikin dan Ibu Hartati.

Pendidikan formal penulis dimulai dari Taman Kanak-Kanak di TK Angelly Prabumulih, kemudian melanjutkan pendidikan dasar di SD Angelly Prabumulih. Selanjutnya, penulis menempuh pendidikan menengah pertama di SMP Negeri 1 Prabumulih dan pendidikan menengah atas di SMA Negeri 1 Prabumulih.

Pada tahun 2022, penulis melanjutkan pendidikan di Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA), Universitas Lampung melalui jalur Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN). Selama masa perkuliahan, penulis memiliki ketertarikan pada bidang matematika dan statistika, khususnya dalam analisis data dan pemrograman.

Selama menempuh pendidikan di perguruan tinggi, penulis aktif dalam kegiatan akademik dan organisasi. Penulis merupakan anggota Bidang Keilmuan Himpunan Mahasiswa Jurusan Matematika (HIMATIKA) Universitas Lampung pada tahun 2023. Selain itu, penulis turut berpartisipasi dalam kompetisi nasional, yaitu Pagelaran Mahasiswa Nasional Bidang Teknologi Informasi dan Komunikasi (GEMASTIK) pada tahun 2024 dan 2025.

Sebagai bentuk pengembangan kompetensi dan penerapan ilmu, penulis melaksanakan Praktik Kerja Lapangan (PKL) di PT Bank Rakyat Indonesia (Persero) Tbk pada Desember 2024 hingga Januari 2025. Penulis juga mengikuti program Magang MBKM Mandiri di Badan Riset dan Inovasi Nasional (BRIN) KST Samaun Samadikun pada Februari hingga Juli 2025.

KATA INSPIRASI

”Janganlah hendaknya kamu kuatir tentang apa pun juga, tetapi nyatakanlah dalam segala hal keinginanmu kepada Allah dalam doa dan permohonan dengan ucapan syukur.”

(Filipi 4:6)

”Janganlah takut, sebab Aku menyertai engkau, janganlah bimbang, sebab Aku ini Allahmu; Aku akan meneguhkan, bahkan akan menolong engkau; Aku akan memegang engkau dengan tangan kanan-Ku yang membawa kemenangan.”

(Yesaya 41:10)

”Mintalah, maka akan diberikan kepadamu; carilah, maka kamu akan mendapat; ketoklah, maka pintu akan dibukakan bagimu.”

(Matius 7:7)

”Serahkanlah perbuatanmu kepada TUHAN, maka terlaksanalah segala rencanamu.”

(Amsal 16:3)

”Karena masa depan sungguh ada, dan harapanmu tidak akan hilang.”

(Amsal 23:18)

PERSEMBAHAN

Dengan penuh rasa syukur kepada Tuhan Yesus Kristus atas kasih, berkat, dan penyertaan-Nya sehingga skripsi ini dapat diselesaikan dengan baik dan tepat pada waktunya. Dengan rasa syukur dan bahagia, penulis mempersembahkan ungkapan terima kasih yang tulus kepada:

Papa dan Mamaku Tercinta

Terima kasih kepada kedua orang tua tercinta atas segala pengorbanan, kasih sayang, doa, dan dukungan yang senantiasa diberikan. Terima kasih atas setiap nasihat, bimbingan, serta pelajaran hidup yang membentuk penulis menjadi pribadi yang lebih kuat dan memahami makna ketulusan serta perjuangan. Semoga penulis dapat menjadi pribadi yang bermanfaat bagi banyak orang dan membawa kebanggaan bagi keluarga.

Dosen Pembimbing dan Pembahas

Terimakasih kepada dosen pembimbing dan pembahas yang sudah sangat membantu, memberikan motivasi, memberikan arahan serta ilmu yang berharga.

Sahabat-sahabatku

Terima kasih kepada semua orang-orang baik yang telah memberikan pengalaman, dukungan, motivasi, serta doa selama proses ini. Setiap kebersamaan menjadi bagian berharga dalam perjalanan penulis. Semoga di masa mendatang dapat bertemu kembali dengan cerita keberhasilan masing-masing.

Almamater Tercinta
Universitas Lampung

SANWACANA

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat, rahmat, dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini yang berjudul "Analisis Parameter Terbaik dalam Deteksi Kesalahan Label I-DIS pada *Named Entity Recognition* Judul Berita Penyakit Menular Berbasis *Hybrid* IndoBERT–BiLSTM–CRF" dengan baik dan lancar serta tepat pada waktu yang telah ditentukan.

Dalam proses penyusunan skripsi ini, banyak pihak yang telah membantu memberikan bimbingan, dukungan, arahan, motivasi serta saran sehingga skripsi ini dapat terselesaikan. Oleh karena itu, dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Ibu Dr. Dian Kurniasari, S.Si.,M.Sc. selaku Pembimbing I yang telah banyak meluangkan waktunya untuk memberikan arahan, bimbingan, motivasi, saran serta dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
2. Ibu Dr. Purnomo Husnul Khotimah, M.T. selaku Pembimbing II yang telah memberikan arahan, bimbingan dan dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
3. Ibu Prof. Dra. Wamiliana, MA.,Ph.D. selaku Penguji yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis sehingga dapat menjadi lebih baik lagi.
4. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
5. Bapak Dr. Ahmad Faisol, S.Si., M.Sc. selaku dosen pembimbing akademik.
6. Seluruh dosen, staff dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Keluarga tercinta, penulis mengucapkan terima kasih yang sebesar-besarnya kepada Papa Gunawan Malikin dan Mama Hartati atas doa, kasih sayang, serta pengorbanan yang tidak pernah terhenti. Dukungan dan kepercayaan yang diberikan menjadi kekuatan utama bagi penulis dalam menyelesaikan skripsi ini. Terima kasih juga kepada adik-adik tersayang, Margaretta Stephanie Gunawan dan Octavia Josephine Gunawan, atas doa dan semangat yang selalu diberikan. Kehadiran kalian menjadi motivasi bagi penulis untuk terus berjuang dan memberikan yang terbaik.
8. Sahabat penulis sejak SD, yaitu Indah Dwi Jaya, yang telah banyak memberikan tawa dan selalu ada di setiap cerita, baik saat penulis senang maupun menghadapi kesulitan. Terima kasih karena tidak pernah lelah meluangkan waktu untuk saling membantu dan menguatkan hingga sampai di titik ini. Terima kasih sudah menjadi tempat pulang dan teman berjuang sejauh ini.
9. Sahabat yang sudah seperti keluarga di perantauan, "Laskar Kristus", yaitu Sabet, Irene, dan Elsi, yang selalu menjadi tempat berbagi keluh kesah selama penulis menjalani masa perantauan, sejak awal menjadi mahasiswa baru hingga sampai di titik ini.
10. Terima kasih kepada teman-teman MBKM BRIN dan teman-teman satu bimbingan yang telah menemani, mendukung, dan menguatkan selama proses penyusunan skripsi. Terima kasih kepada Oja, Erin, dan Fatur yang hadir di kedua perjalanan tersebut dan menjadi teman seperjuangan. Terima kasih juga kepada teman MBKM, Nazla dan Nisa, serta teman-teman satu bimbingan, Nadia, Khusni, Mei, Fadillah, Rizki, dan Benaya, atas kebersamaan, bantuan, dan semangat yang membuat setiap proses terasa lebih ringan.

Semoga skripsi ini dapat bermanfaat bagi kita semua. Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, sehingga penulis mengharapkan kritik dan saran yang membangun untuk menjadikan skripsi ini lebih baik lagi.

Bandar Lampung, 13 April 2026

Anita Caroline Gunawan

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Tingginya intensitas pemanfaatan media digital dalam kehidupan sehari-hari memengaruhi pola masyarakat dalam memperoleh dan menyebarkan informasi. Ketergantungan terhadap platform digital untuk mengakses berita menyebabkan peningkatan signifikan pada volume artikel daring setiap hari. Beberapa portal berita nasional secara rutin mempublikasikan ratusan artikel dari berbagai kategori, termasuk kesehatan. Berdasarkan data Asosiasi Penyelenggara Jasa Internet Indonesia (APJII, 2025), jumlah pengguna internet di Indonesia mencapai 229.428.417 jiwa dari total populasi 284.438.900 jiwa, sehingga menjadikan Indonesia sebagai salah satu negara dengan pengguna internet terbesar di Asia. Tingginya volume dan kecepatan arus informasi tersebut menimbulkan tantangan dalam menemukan serta menyeleksi berita yang relevan di antara kumpulan teks tidak terstruktur. Oleh karena itu, diperlukan sistem komputasional yang mampu mengekstraksi informasi penting secara otomatis agar data dapat dimanfaatkan secara efisien dan akurat.

Salah satu topik berita yang memiliki dampak besar terhadap masyarakat adalah kesehatan, khususnya mengenai penyakit menular. Informasi dalam berita kesehatan daring tidak hanya bersifat informatif, tetapi juga berpotensi memengaruhi perilaku masyarakat dan kebijakan publik. Penyakit seperti COVID-19, tuberkulosis, diare, dan HIV/AIDS masih menjadi perhatian utama dalam kesehatan masyarakat di Indonesia. Artikel daring mengenai penyakit menular umumnya mengandung informasi yang dapat dikategorikan berdasarkan unsur 5W1H (*what, who, when, where, why, how*), sehingga berpotensi diolah menjadi data terstruktur untuk analisis yang lebih mendalam. Informasi seperti nama penyakit, lokasi kejadian, organisasi terkait, dan waktu peristiwa dapat diolah menjadi data kuantitatif untuk mendukung

analisis distribusi, dan tren penyebaran penyakit (Putra & Kurniawan, 2021). Analisis tersebut penting untuk mendukung pemantauan kesehatan masyarakat berbasis data (*data-driven monitoring*) dan pengambilan keputusan berbasis bukti (*evidence-based decision making*).

Dalam konteks tersebut, dibutuhkan pendekatan komputasional yang mampu mengenali pola dan entitas penting dalam teks secara otomatis. Salah satu pendekatan utama dalam *Natural Language Processing* (NLP) adalah *Named Entity Recognition* (NER), yaitu teknik untuk mengidentifikasi dan mengklasifikasikan entitas tertentu dari teks tidak terstruktur, seperti nama orang, organisasi, lokasi, tanggal, dan istilah penting lainnya (Manurung dkk., 2025). Tujuannya adalah mengekstraksi informasi bermakna dari teks mentah agar dapat diubah menjadi data terstruktur yang berguna dalam berbagai aplikasi, seperti sistem tanya jawab otomatis, ekstraksi informasi medis, dan analisis berita kesehatan daring.

Pengembangan NER untuk Bahasa Indonesia menghadapi sejumlah tantangan dibandingkan dengan bahasa lain seperti Inggris atau Mandarin. Keterbatasan data beranotasi, kompleksitas struktur bahasa, dan variasi gaya penulisan dalam berita daring menjadi hambatan utama (Karo dkk., 2025). Bahasa Indonesia memiliki karakter morfologis aglutinatif, penggunaan huruf kapital yang tidak konsisten, serta banyak kata yang bersifat ambigu tergantung pada konteks misalnya kata Corona yang dapat merujuk pada penyakit, lokasi, atau organisasi. Kondisi tersebut menjadikan pengenalan entitas lebih kompleks dan memerlukan fitur linguistik tambahan agar model dapat memahami konteks dengan lebih baik.

Salah satu fitur linguistik penting dalam sistem NER adalah *Part-of-Speech* (POS) *tagging*, yaitu proses pemberian label kategori sintaksis pada setiap kata dalam kalimat. Informasi POS membantu model membedakan entitas yang memiliki bentuk kata serupa tetapi fungsi berbeda. Selain itu, representasi vektor kata (*word embedding*) berperan penting dalam memetakan makna semantik antar kata. Salah satu metode *word embedding* yang banyak digunakan adalah FastText, yaitu model berbasis *neural network* yang merepresentasikan kata sebagai kumpulan *subword*, sehingga mampu menangkap informasi morfologis serta mengatasi permasalahan *out-of-vocabulary* (OOV). Di sisi lain, IndoBERT merupakan model bahasa berbasis arsitektur *Bidirectional Encoder Representations from Transformers* (BERT) yang telah dilatih khusus menggunakan korpus Bahasa Indonesia, sehingga mampu

memahami konteks kata secara mendalam dalam suatu kalimat. Kombinasi FastText dan IndoBERT dapat memperkuat representasi linguistik karena menggabungkan keunggulan informasi morfologis dan pemahaman konteks, sehingga meningkatkan kinerja model dalam tugas NER berbahasa Indonesia.

Berbagai penelitian terdahulu menunjukkan pengembangan NER terus berkembang seiring kemajuan model *deep learning*. Menurut penelitian Jiang dkk. (2022), korpus Tweebank-NER untuk teks media sosial digunakan untuk membandingkan performa beberapa model NLP, dan hasilnya menunjukkan bahwa BERTweet unggul dalam POS *tagging* dan *dependency parsing*, sedangkan Stanza berbasis BiLSTM–CRF terbaik dalam tokenisasi dan *lemmatization*. Penelitian oleh Bhadauria dkk. (2024) menunjukkan bahwa kualitas data berpengaruh signifikan terhadap performa NER model BiLSTM dengan Flair *embedding* mencapai F1-*score* tertinggi 91% pada data bersih, tetapi menurun drastis ketika data mengandung *noise*. Hal ini mendukung pentingnya *embedding* robust seperti FastText untuk menangani variasi ejaan dan kesalahan pengetikan. Menurut Khairunnisa dkk. (2023), IndoBERT monolingual memperoleh F1-*score* tertinggi 91%, melampaui model multibahasa seperti XLM-RoBERTa, sehingga membuktikan kemampuannya memahami konteks Bahasa Indonesia secara efektif.

Selanjutnya, penelitian Wei dkk. (2022) menunjukkan bahwa model RoBERTa–BiLSTM–CRF efektif untuk mengekstraksi entitas teori pada artikel ilmiah, dengan *precision* tertinggi 89,72%, sedangkan penelitian Yulianti dkk. (2024) menemukan bahwa model XLM-RoBERTa memperoleh F1-*score* 0,9295 pada dokumen hukum, lebih tinggi dibanding BiLSTM–CRF (0,85). Kedua penelitian tersebut menegaskan bahwa arsitektur hibrida Transformer–BiLSTM–CRF unggul dalam pengenalan entitas lintas domain. Secara konseptual, CRF memodelkan probabilitas transisi antarlabel, sedangkan BiLSTM mempelajari representasi kata berdasarkan konteks kiri dan kanan, sehingga proses pelabelan menjadi lebih konsisten dan akurat.

Berdasarkan berbagai penelitian tersebut, penerapan *embedding* kontekstual, fitur linguistik POS *tagging*, dan arsitektur hibrida terbukti efektif dalam meningkatkan performa sistem NER. Namun, sebagian besar penelitian masih berfokus pada peningkatan akurasi tanpa disertai analisis kesalahan label (*error analysis*) yang mendalam. Selain itu, penelitian yang secara khusus menerapkan pendekatan

IndoBERT–BiLSTM–CRF dengan POS *tagging* Stanza dan FastText *embedding* pada domain berita penyakit menular masih sangat terbatas. Padahal, judul berita memiliki struktur kalimat yang singkat dan minim konteks, sehingga berpotensi menimbulkan kesalahan pelabelan entitas, terutama pada label sekuensial seperti I-DIS (*Inside-Disease*) dan B-DIS (*Beginning-Disease*). Kesalahan pada label I-DIS sering muncul akibat ketidakkonsistenan model dalam mengenali batas entitas atau kurangnya konteks linguistik pada teks pendek, sehingga perlu dilakukan analisis khusus untuk mengidentifikasi pola kesalahan tersebut.

Dalam skema pelabelan BIO, label I-DIS (*Inside-Disease*) memiliki peran yang krusial dalam menjaga keutuhan representasi entitas penyakit yang umumnya terdiri dari lebih dari satu kata. Berbeda dengan label B-DIS yang hanya menandai awal entitas, I-DIS berfungsi untuk memastikan bahwa seluruh rangkaian kata dalam suatu nama penyakit dikenali sebagai satu kesatuan yang utuh. Ketepatan pelabelan I-DIS menjadi penting karena kesalahan pada bagian ini dapat menyebabkan fragmentasi entitas, sehingga informasi penyakit tidak lagi direpresentasikan secara lengkap dan akurat. Hal ini menunjukkan bahwa konsistensi pelabelan pada bagian dalam entitas memiliki pengaruh langsung terhadap kualitas hasil ekstraksi informasi.

Selain itu, pentingnya label I-DIS dalam penelitian ini berkaitan erat dengan kebutuhan analisis penyakit menular berbasis data yang akurat dan terstruktur. Identifikasi entitas penyakit secara utuh diperlukan untuk mendukung pemetaan distribusi, analisis tren, serta pemantauan penyebaran penyakit secara akurat. Kesalahan pada label I-DIS tidak hanya berdampak pada entitas yang sudah dikenal, tetapi juga dapat menghambat proses identifikasi dan ekstraksi entitas penyakit yang muncul dalam teks, terutama pada variasi penulisan atau istilah yang jarang muncul dalam data pelatihan. Oleh karena itu, analisis khusus terhadap kesalahan label I-DIS menjadi penting untuk meningkatkan keandalan model dalam mengekstraksi informasi penyakit secara komprehensif.

Dalam upaya meningkatkan kemampuan model dalam mengatasi kesalahan pelabelan entitas, khususnya pada label I-DIS, performa model NER sangat dipengaruhi oleh kombinasi parameter pelatihan seperti *learning rate*, *batch size*, jumlah *epoch*, dan ukuran *hidden layer*. Pemilihan parameter yang tepat berpengaruh langsung terhadap konvergensi model, stabilitas pelatihan, dan akurasi akhir. Oleh karena itu, analisis parameter terbaik (*hyperparameter tuning*) menjadi langkah

penting untuk memastikan bahwa model tidak hanya akurat, tetapi juga efisien dan adaptif terhadap data baru.

Untuk menjawab kesenjangan tersebut, penelitian ini mengusulkan pendekatan hibrida IndoBERT–BiLSTM–CRF yang menggabungkan keunggulan ketiga model. IndoBERT digunakan untuk menangkap representasi kontekstual kata, BiLSTM untuk mempelajari pola urutan kata secara dua arah, dan CRF untuk memodelkan transisi antarlabel agar hasil prediksi lebih konsisten. Fitur tambahan berupa POS *tagging* Stanza dan FastText *embedding* digunakan untuk mengatasi ambiguitas linguistik dan masalah OOV pada teks Bahasa Indonesia. Penelitian ini juga melakukan *error analysis* secara statistik untuk menghitung frekuensi dan distribusi kesalahan per label, serta mengevaluasi performa model menggunakan metrik *precision*, *recall*, dan F1-*score*.

Dengan pendekatan tersebut, penelitian ini diharapkan memberikan kontribusi empiris terhadap pengembangan NER Bahasa Indonesia yang lebih akurat dan adaptif, serta memperkuat integrasi antara pembelajaran mesin dan analisis statistik. Hasil penelitian juga diharapkan bermanfaat bagi analisis data kesehatan masyarakat, khususnya dalam pemantauan dan deteksi penyebaran penyakit menular berbasis berita daring.

1.2 Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini yaitu:

1. Bagaimana membangun model *Named Entity Recognition* (NER) berbasis arsitektur *hybrid* IndoBERT–BiLSTM–CRF dengan tambahan fitur POS *tagging* Stanza dan *embedding* FastText untuk pemrosesan judul berita penyakit menular?
2. Bagaimana menentukan kombinasi parameter pelatihan terbaik (*hyperparameter tunning*) untuk mengoptimalkan kinerja model NER pada judul berita penyakit menular?
3. Bagaimana menganalisis pola kesalahan pelabelan, khususnya pada label I-DIS (*Inside-Disease*), untuk mengidentifikasi sumber ketidakakuratan dan mendukung perbaikan model NER?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini yaitu:

1. Membangun model *Named Entity Recognition* (NER) berbasis arsitektur *hybrid* IndoBERT–BiLSTM–CRF dengan fitur POS *tagging* Stanza dan *embedding* FastText untuk pemrosesan judul berita penyakit menular.
2. Menentukan parameter pelatihan terbaik untuk mengoptimalkan kinerja model NER pada judul berita penyakit menular.
3. Menganalisis pola kesalahan pelabelan, khususnya pada label I-DIS, guna mengidentifikasi sumber ketidakakuratan dan meningkatkan kualitas prediksi model.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini yaitu:

1. Menghasilkan model NER yang lebih akurat dan konsisten untuk ekstraksi entitas pada judul berita penyakit menular berbahasa Indonesia.
2. Memberikan pemahaman empiris terkait pengaruh fitur POS Stanza dan *embedding* FastText terhadap performa model NER.
3. Menyediakan informasi pola kesalahan pelabelan, khususnya label I-DIS, sebagai dasar perbaikan model dan referensi penelitian selanjutnya.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Sub bagian ini menguraikan beberapa penelitian yang telah dilakukan oleh peneliti sebelumya dan digunakan sebagai referensi untuk penelitian ini. Ringkasan penelitian terdahulu disajikan dalam Tabel 1 sebagai berikut:

Tabel 1. Penelitian Terdahulu.

No.	Penelitian	Data	Metode	Hasil (%)		
				Prec	Rec	F1
1.	*Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis* (Jiang dkk., 2022)	Domain: Bahasa Inggris (Twitter) Jumlah: 3.550 *tweet* Sumber: Tweebank V2	*Tagging*: NER (PER, ORG, LOC, MISC) Model: BERTweet, XLM-RoBERTa, Stanza (BiLSTM–CRF), FLAIR, spaCy	–	–	74.35
2.	*The Effects of Data Quality on Named Entity Recognition* (Bhadauria dkk., 2024)	Domain: Bahasa Inggris *Dataset*: CoNLL 2003, WNUT 16, OntoNotes v5	*Tagging*: NER (BIO) Model: CRF, BERT, BiLSTM–Flair	–	–	85.00
3.	*Dataset Enhancement & Multilingual Transfer for NER in Indonesian* (Khairunnisa dkk., 2023)	Domain: Bahasa Indonesia *Dataset*: S&N (2016, *re-annotation*) Sumber: Korpus berita	*Tagging*: NER (BIO) Model: BiLSTM–CRF, IndoBERT, IndoLEM, mBERT, XLM-RoBERTa	94.93	94.87	94.90

No.	Penelitian	Data	Metode	Hasil (%)		
				Prec	Rec	F1
4.	*Theory Entity Extraction for Social and Behavioral Sciences Papers* (Wei dkk., 2022)	Domain: Bahasa Inggris Sumber: DARPA *SCORE*	*Tagging*: NER (BIO) Model: RoBERTa–BiLSTM–CRF, BiLSTM, Transformer, GCN *Embedding*: FastText, GloVe, ELMo, BERT, GPT, RoBERTa	89.72	67.76	77.21
5.	*NER on Indonesian Legal Documents* (Yulianti dkk., 2024)	Domain: Bahasa Indonesia Sumber: Mahkamah Agung RI	*Tagging*: NER (20 label entitas hukum) Model: XLM-RoBERTa, IndoBERT, IndoRoBERTa, BiLSTM–CRF	91.24	94.72	92.95

2.1.1 Penelitian Pertama (Jiang dkk., 2022)

Penelitian ini berfokus pada pembangunan korpus dan pengembangan model *Natural Language Processing* (NLP) untuk analisis media sosial, khususnya Twitter yang memiliki karakteristik bahasa tidak baku dan struktur kalimat kompleks. *Dataset* yang digunakan bernama Tweebank-NER, hasil anotasi entitas bernama pada korpus Tweebank V2 dengan kerangka *Universal Dependencies* (UD). *Dataset* ini terdiri atas 3.550 *tweet* anonim berbahasa Inggris dengan pembagian 24.000 *token* data latih, 11.000 *token* validasi, dan 19.000 *token* uji. Anotasi dilakukan secara *crowdsourcing* melalui *Amazon Mechanical Turk* dengan minimal tiga anotator per *tweet*, mengikuti panduan CoNLL 2003 yang mencakup empat kelas utama: *person* (PER), *organization* (ORG), *location* (LOC), dan *miscellaneous* (MISC). Tingkat kesepakatan antar anotator diukur menggunakan *pairwise* F1-*score* tanpa label "O", dengan rata-rata keseluruhan sebesar 70,7%.

Pada tahap pemodelan, dikembangkan pipeline NLP bernama Twitter-Stanza yang mencakup tokenisasi, *lemmatization*, POS *tagging*, *dependency parsing*, dan NER. Model yang diuji meliputi Stanza, spaCy, FLAIR, serta Transformer seperti BERTweet dan XLM-RoBERTa. Untuk tugas NER, Stanza dilatih menggunakan arsitektur BiLSTM–CRF dengan GloVe *embedding* berdimensi 100 yang diadaptasi untuk bahasa Twitter. Hasil evaluasi menunjukkan bahwa BERTweet mencapai performa terbaik dengan F1-*score* 74,35%, sedangkan Stanza memperoleh 62,53% dengan ukuran model 75% lebih kecil dibanding FLAIR. Kesalahan terbanyak terjadi pada entitas MISC yang sering diklasifikasikan sebagai "O". Penelitian ini menegaskan efektivitas model Transformer dalam menangani teks informal dan berisik (*noisy text*), seperti halnya berita daring yang menjadi fokus penelitian ini.

2.1.2 Penelitian Kedua (Bhadauria dkk., 2024)

Penelitian ini menyoroti pengaruh kualitas data terhadap performa model *Named Entity Recognition* (NER) dengan menganalisis dampak *noise* atau gangguan data terhadap hasil prediksi. Empat jenis *noise* yang dikaji meliputi kesalahan ejaan, pengetikan, pengenalan karakter optik (*Optical Character Recognition error*), dan pemendekan kalimat (*Sentence Shortening Error*). Tiga model dibandingkan, yaitu *Conditional Random Field* (CRF), BERT, dan BiLSTM dengan Flair *embedding*, menggunakan tiga dataset populer beranotasi skema BIO: CoNLL 2003, WNUT 16, dan OntoNotes v5.

Hasil menunjukkan bahwa peningkatan *noise* secara signifikan menurunkan nilai F1-*score* pada seluruh model. Penurunan terbesar terjadi pada BERT saat diuji dengan *composite noise*, sedangkan BiLSTM–Flair menunjukkan ketahanan lebih baik terhadap gangguan data, dan CRF cenderung stabil pada kesalahan tunggal namun menurun tajam ketika *noise* dikombinasikan. Pada dataset WNUT 16, penurunan F1-*score* terbesar mencapai 0,26 poin pada BERT, sedangkan pada OntoNotes v5, kesalahan ejaan menurunkan skor BiLSTM–Flair dari 0,85 menjadi 0,23. Penelitian ini menegaskan pentingnya kualitas data serta tahapan pembersihan dan validasi dalam membangun sistem ekstraksi entitas yang tangguh terhadap *noise*, sekaligus menyoroti peran *error analysis* untuk memahami sumber kesalahan prediksi model NER.

2.1.3 Penelitian Ketiga (Khairunnisa dkk., 2023)

Penelitian ini berfokus pada peningkatan kualitas *Named Entity Recognition* (NER) berbahasa Indonesia melalui proses *re-annotation* dataset serta penerapan pendekatan *monolingual* dan *cross-lingual transfer learning*. *Dataset* S&N (2016) memiliki ketidakkonsistenan anotasi hingga 30%, terutama pada entitas *organization*, sehingga dilakukan *re-annotation* untuk menghasilkan data yang lebih bersih dan konsisten. Proses ini meningkatkan performa model secara signifikan, dengan F1-*score* 90,85 menggunakan BiLSTM–CRF dan FastText *embedding*, yang menunjukkan pentingnya kualitas anotasi terhadap keandalan model. Penelitian juga membandingkan beberapa pendekatan, meliputi BiLSTM–CRF, *Transformer fine-tuning* (BERT), serta kombinasi BiLSTM–CRF dengan berbagai *embedding* seperti IndoBERT, IndoLEM, mBERT, dan XLM-RoBERTa. Hasil menunjukkan bahwa kombinasi BiLSTM–CRF dengan IndoBERT memberikan kinerja terbaik dengan F1-*score* 94,90% karena kesesuaian kosakata IndoBERT dengan korpus berita Indonesia dan kemampuan BiLSTM–CRF menangkap pola sekuensial secara lebih efektif dibanding *fine-tuning* murni.

Selain itu, penelitian ini mengeksplorasi *unsupervised cross-lingual transfer learning* dari empat bahasa sumber daya tinggi (Inggris, Spanyol, Belanda, dan Jerman) menggunakan strategi *vector-based transfer*, NMT-*based alignment*, parallel corpora, dan *teacher–student model*. Pendekatan *vector-based transfer* menunjukkan hasil terbaik untuk *single-source* maupun *multi-source*, dengan Belanda dan Inggris sebagai kontributor utama karena kemiripan leksikal dengan bahasa Indonesia. Penelitian ini menegaskan efektivitas kombinasi IndoBERT–BiLSTM–CRF sebagai pendekatan hibrida yang relevan untuk meningkatkan kinerja NER, sebagaimana diadaptasi pula dalam penelitian ini.

2.1.4 Penelitian Keempat (Wei dkk., 2022)

Penelitian ini membahas ekstraksi entitas teori (*theory entity extraction*) pada makalah ilmiah bidang ilmu sosial dan perilaku (*Social and Behavioral Sciences* – SBS) menggunakan pendekatan *distant supervision*. Latar belakang penelitian ini adalah keterbatasan data beranotasi untuk tugas ekstraksi teori, yang menghambat penerapan metode *supervised learning*. Untuk mengatasinya, dikembangkan

kerangka otomatis yang memanfaatkan Wikipedia sebagai sumber pengetahuan guna menghasilkan korpus beranotasi secara otomatis. Dari 2.400 makalah SBS dalam proyek DARPA *Systematizing Confidence in Open Research and Evidence* (SCORE), dihasilkan sekitar 870.000 kalimat dengan 550 entitas teori unik. Anotasi dilakukan secara otomatis melalui web scraping teori dari Wikipedia, konversi dokumen menggunakan GROBID, segmentasi kalimat dengan Stanza, serta *indexing* berbasis Elasticsearch untuk pelabelan entitas dalam skema BIO. Pendekatan ini mampu menghasilkan data beranotasi hanya dalam dua jam, jauh lebih efisien dibandingkan anotasi manual.

Empat arsitektur *deep learning* dibandingkan, yaitu BiLSTM, BiLSTM–CRF, Transformer, dan *Graph Convolutional Network* (GCN), dengan berbagai *embedding* seperti FastText, GloVe, ELMo, BERT, GPT, dan RoBERTa. Hasil eksperimen menunjukkan bahwa model RoBERTa–BiLSTM–CRF memberikan performa terbaik dengan *precision* 89,72%, *recall* 67,76%, dan F1-*score* 77,21%. Pendekatan *distant supervision* terbukti efektif dalam mengatasi keterbatasan data beranotasi dan dapat diadaptasi untuk domain lain. Penelitian ini relevan karena menerapkan arsitektur hibrida Transformer–BiLSTM–CRF yang serupa dengan pendekatan yang digunakan dalam penelitian ini.

2.1.5 Penelitian Kelima (Yulianti dkk., 2024)

Penelitian ini menerapkan *Named Entity Recognition* (NER) pada dokumen hukum berbahasa Indonesia untuk mengekstraksi informasi penting dari putusan pengadilan. *Dataset* yang digunakan, *Indonesian Legal Entity Recognition* (IndoLER), terdiri atas 1.000 dokumen putusan pidana dari Mahkamah Agung Republik Indonesia dengan total sekitar enam juta kata dan lebih dari 25.000 entitas teranotasi. Anotasi dilakukan secara manual menggunakan skema BIO dengan 20 kategori entitas hukum, seperti nama hakim, pasal dakwaan, dan amar putusan. Nilai kesepakatan antar anotator (*Fleiss' Kappa*) mencapai 0,9519, menunjukkan konsistensi tinggi. *Dataset* kemudian dibagi menjadi 70% data latih, 20% data uji, dan 10% data validasi menggunakan metode *stratified shuffle split*.

Penelitian ini membandingkan beberapa model Transformer yaitu, M-BERT, IndoBERT, IndoRoBERTa, dan XLM-RoBERTa serta dua model *baseline*, yaitu

BiLSTM dan BiLSTM–CRF. Hasil menunjukkan bahwa XLM-RoBERTa (*large*) mencapai performa terbaik dengan *precision* 91,24%, *recall* 94,72%, dan F1-*score* 92,95%, melampaui BiLSTM dan BiLSTM–CRF masing-masing sebesar 7,9% dan 2,6%. Analisis kesalahan menunjukkan bahwa misklasifikasi terbanyak berasal dari label "O" akibat ketidakseimbangan data. Penelitian ini menegaskan efektivitas Transformer dalam tugas NER berbahasa Indonesia serta pentingnya *error analysis* untuk memahami distribusi kesalahan label dalam model.

2.2 Artikel Berita Online

Artikel berita online merupakan perkembangan media berita konvensional yang sebelumnya dipublikasikan melalui koran atau majalah dan kini dipublikasikan melalui platform digital berbasis internet (Westlund dkk., 2025). Informasi disajikan dalam bentuk teks, gambar, audio, maupun video, menjadikannya variatif, interaktif, dan mudah diakses melalui situs web berita, portal jurnalistik digital, serta media sosial. Prinsip kecepatan, akurasi, dan interaktivitas yang diterapkan media daring memungkinkan distribusi informasi secara *real-time* dengan jangkauan luas (Achonwa & Adedeji, 2025). Perkembangan ini memperkuat posisi berita online sebagai sumber informasi utama masyarakat modern yang menyesuaikan diri dengan pola konsumsi digital.

Menurut Williams & Archibong (2024), perkembangan berita online tidak terlepas dari fungsi media massa yang mencakup penyebaran informasi, penyampaian interpretasi, pembentukan kesepakatan, penghubung sosial, dan transmisi nilai budaya. Fungsi tersebut dijalankan melalui publikasi yang responsif terhadap dinamika peristiwa, sehingga pembaruan informasi dapat dilakukan secara cepat dan terstruktur. Dalam konteks penyampaian informasi, judul berita berperan penting karena memberikan gambaran awal mengenai isi konten dan menonjolkan unsur "siapa" serta "apa yang dilakukan" melalui pemilihan kata yang padat dan langsung. Tradisi penonjolan judul dari media cetak tetap dipertahankan dalam format digital melalui penggunaan tipografi dan tata visual yang menarik agar pembaca terarah pada topik utama. Dengan karakteristik tersebut, berita online menjadi sumber informasi yang tidak hanya dinamis, tetapi juga menarik untuk dikaji dalam analisis teks dan linguistik.

2.3 Penyakit Menular

Penyakit menular merupakan penyakit yang berpindah dari satu individu ke individu lain melalui penularan langsung maupun melalui perantara (Ryu dkk., 2022). Mikroorganisme seperti virus, bakteri, jamur, dan parasit menjadi penyebab utama infeksi yang masuk ke dalam tubuh manusia dan menimbulkan gangguan kesehatan. Mekanisme penularannya mencakup kontak fisik, percikan pernapasan, udara, serta makanan, minuman, atau benda yang terkontaminasi, termasuk penularan melalui vektor seperti nyamuk (Zhou dkk., 2024). Variasi jalur penularan tersebut menunjukkan bahwa penyakit menular memiliki tingkat penyebaran yang dinamis serta dapat berkembang secara cepat pada lingkungan dengan mobilitas tinggi.

Contoh penyakit menular yang sering ditemui di Indonesia mencakup COVID-19, Tuberkulosis (TBC), dan Demam Berdarah *Dengue* (DBD) yang memiliki potensi penyebaran cepat dan berdampak luas terhadap kesehatan masyarakat. Penyakit tersebut menjadi perhatian dalam kajian kesehatan publik karena mempengaruhi pola interaksi sosial, kapasitas layanan kesehatan, dan persepsi risiko masyarakat. Penyakit menular memiliki karakteristik penyebaran yang dipengaruhi oleh faktor lingkungan, perilaku masyarakat, serta kemampuan patogen beradaptasi, sehingga kajiannya penting untuk memahami dinamika kesehatan publik.

2.4 *Natural Language Processing* (NLP)

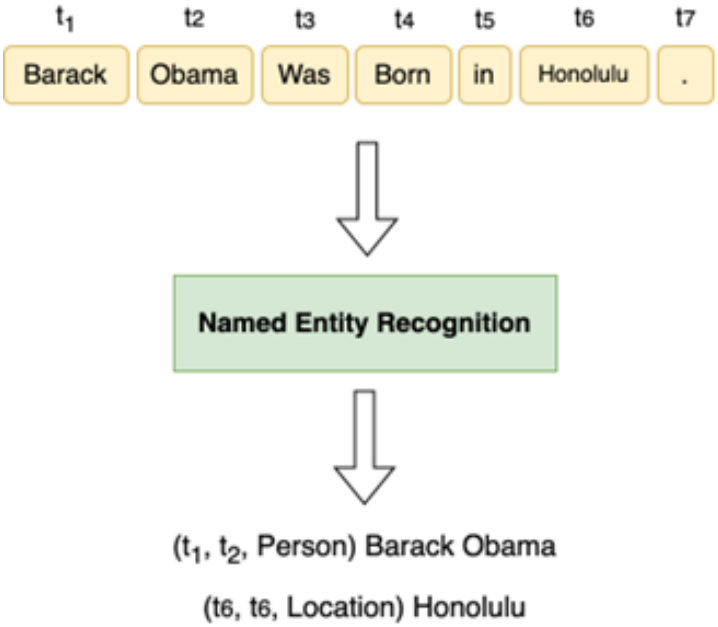
Natural Language Processing (NLP) merupakan cabang kecerdasan buatan yang mengembangkan pendekatan komputasional agar sistem dapat memahami, memproses, dan menghasilkan bahasa manusia secara terstruktur (Puspitasari dkk., 2024). NLP berlandaskan integrasi kecerdasan buatan, linguistik komputasional, dan representasi numerik sehingga mesin mampu menafsirkan makna bahasa serta memberikan respons sesuai konteks. NLP menggabungkan teori linguistik, statistika, dan teknik komputasi untuk mengonversi data teks tidak terstruktur menjadi informasi yang dapat dianalisis secara sistematis (Hou & Huang, 2025). Pendekatan ini membentuk kerangka analitik yang memungkinkan sistem mengidentifikasi pola bahasa, memetakan struktur kalimat, dan mengekstraksi makna secara konsisten.

Ruang lingkup NLP mencakup tokenisasi, analisis sintaksis, pemodelan bahasa, ekstraksi informasi, dan klasifikasi teks yang menyediakan landasan bagi pemahaman bahasa secara komputasional. Perkembangan arsitektur *deep learning*, terutama model transformer, telah meningkatkan kemampuan sistem dalam memahami konteks kalimat dan meningkatkan ketepatan analisis semantik. Penerapan NLP pada Bahasa Indonesia memiliki nilai strategis karena mampu menangani variasi morfologi, struktur kalimat, serta kompleksitas teks daring. Dengan demikian, NLP menjadi dasar pengembangan metode identifikasi entitas, analisis dokumen, dan sistem pengolahan berita digital yang memerlukan presisi interpretasi terhadap makna linguistik.

2.5 *Named Entity Recognition* (NER)

Named Entity Recognition (NER) merupakan tahap fundamental dalam ekstraksi informasi pada pemrosesan bahasa alami yang berfungsi mengidentifikasi serta mengklasifikasikan entitas penting dalam teks, seperti nama orang, lokasi, organisasi, tanggal, waktu, dan ekspresi numerik (Pakhale, 2023). Proses penandaan dilakukan melalui analisis *token* berdasarkan makna dan konteks untuk memperoleh informasi relevan dari dokumen terstruktur maupun tidak terstruktur. Konsep NER berakar pada gagasan *named entities* yang dikembangkan dalam *Message Understanding Conference* (MUC) untuk mendukung tujuan ekstraksi informasi pada teks tidak terstruktur (Warto dkk., 2024). Perkembangannya mendorong penguatan pendekatan analitis agar sistem mampu mengenali entitas secara konsisten pada berbagai variasi penulisan dan struktur kalimat.

Penerapan NER meluas pada berbagai bidang seperti pengindeksan informasi, sistem tanya jawab, penerjemahan mesin, peringkasan otomatis, dan pemetaan konten digital yang memerlukan identifikasi entitas dengan ketepatan tinggi (Keerthana & Uddin, 2024). Tantangan implementasinya meliputi variasi bentuk penulisan entitas, ketidakteraturan struktur kalimat, dan ambiguitas makna sehingga pengembangan metode berbasis *machine learning*, dan *deep learning* diperlukan untuk meningkatkan akurasi (Keraghel dkk., 2024). Peningkatan teknik pemodelan tersebut memperluas kemampuan sistem dalam mengenali entitas seperti nama, lokasi, tanggal, dan waktu pada korpus berukuran besar. Ilustrasi proses penandaan entitas ditunjukkan pada Gambar 1.



Gambar 1. Contoh Penerapan NER (Keraghel dkk., 2024).

Salah satu skema anotasi yang umum digunakan dalam NER adalah skema BIO (*Begin*, *Inside*, *Outside*) yang menandai posisi setiap token dalam entitas. Skema pelabelan ini terdiri atas tiga kategori, yaitu *Beginning of Named Entity* (B) yang digunakan untuk menandai awal entitas bernama, *Inside of Named Entity* (I) yang menandai bagian lanjutan entitas bernama yang terdiri atas lebih dari satu kata, dan *Outside* (O) yang menandai token yang tidak termasuk dalam entitas (Jarrar dkk., 2022). Penerapan skema BIO menjaga konsistensi hasil anotasi serta memfasilitasi model dalam mempelajari pola urutan kata dan batas entitas secara sistematis sehingga mendukung peningkatan akurasi identifikasi entitas pada tahap pelatihan. Ilustrasi umum proses penandaan dengan skema BIO ditunjukkan pada Gambar 2.

Gambar 2. Contoh Penandaan Entitas dengan skema BIO (Maurya, 2023).

Gambar 2 memperlihatkan contoh bagaimana entitas “Albert” diberi label B-PER sebagai penanda awal entitas orang (*Begin–Person*). Selanjutnya, frasa “United States of America” ditandai berurutan dengan label B-LOC, I-LOC, dan B-LOC, di mana B-LOC menunjukkan awal entitas lokasi (*Begin–Location*), sedangkan I-LOC menandai kata yang merupakan kelanjutan dari entitas lokasi (*Inside–Location*). Sementara itu, kata-kata lain seperti *is going to* tidak termasuk dalam entitas tertentu dan diberi label O (*Outside*). Melalui pola pelabelan seperti ini, sistem dapat mengenali batas entitas dengan tepat serta membedakan kata yang merupakan bagian dari entitas dan yang bukan. Dengan demikian, pendekatan NER menjadi komponen penting dalam sistem ekstraksi informasi, termasuk pada analisis teks berita yang berkaitan dengan isu kesehatan atau penyakit menular.

2.6 *Preprocessing*

Preprocessing merupakan tahap fundamental dalam pemrosesan bahasa alami yang menyiapkan teks mentah agar dapat diolah secara optimal oleh sistem komputasional. Proses ini meningkatkan kualitas data melalui penghapusan elemen yang tidak relevan, penyeragaman format penulisan, dan pengaturan struktur teks sehingga informasi linguistik dapat diproses secara akurat (Sujadi, 2022). Dalam konteks Bahasa Indonesia yang memiliki karakteristik khas seperti afiksasi, kata majemuk, dan variasi penulisan, setiap tahap *preprocessing* perlu disesuaikan agar tidak menghilangkan informasi penting, khususnya dalam identifikasi entitas. Secara umum, *preprocessing* mencakup tiga tahap utama berikut.

1. *Cleaning* (Pembersihan Teks) merupakan proses penghapusan elemen yang tidak diperlukan, seperti tanda baca tertentu, karakter non-alfabet, atau simbol yang tidak berkontribusi pada makna kalimat (Lamprou dkk., 2025). Pada teks berita, pembersihan difokuskan pada penghilangan tanda baca yang berpotensi mengganggu pemisahan *token*, tanpa menghapus elemen yang masih relevan.
2. *Normalization* (Normalisasi Teks) bertujuan menyeragamkan bentuk penulisan agar konsisten dan mudah diproses (Finansyah dkk., 2022). Langkah ini mencakup penyeragaman huruf, penyederhanaan variasi penulisan, serta penyesuaian bentuk kata yang berpola serupa untuk mengurangi ketidakkonsistenan data.

3. *Tokenization* (Tokenisasi) adalah proses memecah teks menjadi unit-unit kecil yang disebut *token*, seperti kata atau frasa, yang memiliki makna dalam konteks kalimat (Turuta & Maksymenko, 2025). Proses ini memungkinkan analisis *token* secara independen sehingga pola linguistik dapat dikenali lebih rinci dan digunakan dalam representasi fitur maupun pemodelan.

Penerapan ketiga tahap tersebut menghasilkan teks yang lebih bersih, konsisten, dan terstruktur, sehingga data siap diubah menjadi representasi numerik dan diolah oleh model pembelajaran mesin. *Preprocessing* yang tepat berperan penting dalam meningkatkan kualitas analisis serta meminimalkan kesalahan pada tahap komputasi lanjutan, mendukung performa model secara keseluruhan.

2.7 *Splitting Data*

Splitting data merupakan proses pembagian *dataset* menjadi beberapa subset untuk pelatihan, validasi, dan pengujian model. Menurut Muraina (2022), *splitting* data bertujuan menghindari *overfitting* serta bias dalam pemilihan model dengan memastikan bahwa evaluasi dilakukan pada data yang tidak digunakan selama pelatihan. Pembagian *dataset* dilakukan berdasarkan proporsi tertentu agar distribusi data tetap konsisten dan representatif terhadap keseluruhan *dataset*.

Proses pemisahan data juga berperan dalam meningkatkan kemampuan generalisasi model, karena model yang dilatih dan diuji pada subset berbeda dapat mengenali pola data yang lebih beragam (Kahloot & Ekler, 2021). Tahap ini membantu mencegah bias, menjaga stabilitas performa, serta memastikan model dapat diterapkan secara efektif pada data baru. Dengan demikian, *splitting data* menjadi tahap penting yang menjadikan proses pelatihan, validasi, dan pengujian model berlangsung secara sistematis dan andal.

2.8 *Part of Speech* (POS) *Tagging*

Part-of-Speech (POS) *tagging* merupakan proses dalam pemrosesan bahasa alami yang bertujuan memberikan label kelas kata pada setiap *token* dalam teks (Chiche & Yitagesu, 2022). Label tersebut menunjukkan fungsi tata bahasa kata dalam kalimat

sehingga mendukung analisis linguistik dan pemodelan teks secara sistematis. POS *tagging* berperan penting dalam berbagai aplikasi NLP, seperti ekstraksi informasi, analisis sentimen, dan penerjemahan otomatis.

Stanza merupakan toolkit *open-source* berbasis Python yang mendukung POS *tagging* pada 66 bahasa, termasuk bahasa Indonesia (Qi dkk., 2020). Stanza menggunakan dua tingkatan pelabelan, yaitu *Universal* POS *tags* (UPOS) dan *Language-specific* POS *tags* (XPOS), untuk menghasilkan identifikasi kelas kata yang akurat dan kontekstual sesuai karakteristik bahasa.

1. *Universal* POS *Tags* (UPOS)

UPOS memberikan label kelas kata umum yang berlaku lintas bahasa berdasarkan standar *Universal Dependencies* (UD). Sistem ini terdiri atas 17 tag yang digunakan secara luas dalam pemrosesan bahasa alami. Tabel 2 menampilkan daftar dan deskripsi dari 17 tag UPOS beserta contohnya.

Tabel 2. Daftar 17 Tag UPOS dan Deskripsinya.

Tag	Kategori	Deskripsi	Contoh
ADJ	*Adjective*	Kata sifat yang memodifikasi nomina	tinggi, besar, rentan
ADP	*Adposition*	Preposisi atau postposisi yang menunjukkan hubungan antar kata	di, ke, dari, karena
ADV	*Adverb*	Kata keterangan yang menjelaskan verba atau adjektiva	sangat, tidak, cukup
AUX	*Auxiliary*	Kata kerja bantu yang mendukung verba utama	telah, akan, sedang
CCONJ	*Coordinating Conjunction*	Kata hubung koordinatif	dan, atau, tetapi
DET	*Determiner*	Penentu atau pembatas nomina	ini, itu, para
INTJ	*Interjection*	Kata seru yang mengekspresikan emosi	wah, aduh
NOUN	*Noun*	Nomina umum yang menyatakan benda, konsep, atau ide	penyakit, virus, laporan
NUM	*Numeral*	Kata bilangan	10, 2024, dua
PART	*Particle*	Partikel gramatikal	lah, kah, pun
PRON	*Pronoun*	Kata ganti	ia, mereka, yang
PROPN	*Proper Noun*	Nama diri atau entitas khusus	Indonesia, BRI, Jakarta
PUNCT	*Punctuation*	Tanda baca	., ,
SCONJ	*Subordinating Conjunction*	Kata hubung subordinatif	karena, jika, meskipun

Tag	Kategori	Deskripsi	Contoh
SYM	*Symbol*	Simbol nonalfabetik	%, $
VERB	*Verb*	Kata kerja yang menyatakan tindakan atau proses	dirawat, mengandung, dilaporkan
X	*Other*	Token yang tidak termasuk kategori lain	—

UPOS menyediakan kerangka universal yang menjamin konsistensi pelabelan lintas bahasa dan mendukung interoperabilitas antar model linguistik.

2. *Language-Specific* POS *Tags* (XPOS)

XPOS menyediakan pelabelan yang lebih rinci sesuai karakteristik setiap bahasa. Sistem ini memperluas kategori UPOS dengan menambahkan dimensi morfologis yang lebih spesifik, seperti pembedaan bentuk tunggal dan jamak, bentuk dasar dan lampau, serta variasi gramatikal lainnya.

Jumlah tag XPOS berbeda untuk setiap bahasa. Pada Bahasa Indonesia, Stanza mendefinisikan sekitar 40 tag yang mencakup rincian morfologis berbagai kelas kata. Contoh yang umum digunakan antara lain NN untuk kata benda tunggal, NNS untuk kata benda jamak, VB untuk kata kerja bentuk dasar, VBD untuk kata kerja bentuk lampau, dan JJ untuk kata sifat. Sistem XPOS digunakan dalam analisis linguistik yang memerlukan ketelitian terhadap bentuk morfologis dan struktur sintaksis kalimat.

Kombinasi antara UPOS dan XPOS memungkinkan analisis kelas kata yang komprehensif, di mana UPOS berfungsi memberikan kerangka universal, sedangkan XPOS menambah kedalaman deskriptif terhadap struktur bahasa tertentu.

2.9 FastText *Embedding*

Word embedding merupakan metode representasi kata dalam bentuk vektor bilangan real berdimensi tetap yang dirancang untuk menangkap informasi sintaktis dan semantik secara terstruktur. Representasi tersebut memetakan kata ke dalam ruang vektor laten sehingga kata dengan konteks serupa memiliki jarak vektor yang

berdekatan (Wszola dkk., 2021). Pendekatan ini meningkatkan kinerja model dalam tugas pemrosesan bahasa alami karena memungkinkan sistem mengenali keterkaitan antarkata secara komputasional melalui pola distribusi yang konsisten.

Beragam pendekatan telah dikembangkan untuk menghasilkan representasi kata, antara lain Word2Vec, GloVe, dan FastText. Word2Vec membangun vektor kata berdasarkan distribusi konteks dalam rentang kata di sekitar kata target, sedangkan GloVe memanfaatkan statistik global korpus untuk memperoleh representasi yang stabil. FastText dikembangkan dari Word2Vec mengintegrasikan informasi *subword* berupa karakter n-gram, sehingga representasi kata lebih adaptif terhadap variasi morfologis dan tetap konsisten meskipun kata jarang muncul atau tidak tercantum dalam kosakata pelatihan (*out-of-vocabulary*).

Mekanisme FastText memodelkan kata sebagai himpunan n-gram karakter yang digabung untuk membentuk representasi kata secara utuh. Pendekatan ini mendukung kerangka *Continuous Bag of Words* (CBOW) dan *Skip-gram* untuk menangkap hubungan konteks lokal (Kurniasari dkk., 2025). Kemampuan menghasilkan representasi untuk kata OOV menjadikan FastText lebih adaptif dibanding metode berbasis kata tunggal. Implementasi FastText dalam pemrosesan bahasa alami memberikan keunggulan pada berbagai tugas, seperti klasifikasi teks, ekstraksi entitas, dan analisis konteks, karena mampu memodelkan struktur kata serta variasi morfologis secara mendalam dan efisien.

2.10 *Hyperparameter Tunning*

Hyperparameter tuning merupakan proses pengaturan nilai parameter yang mengendalikan mekanisme pelatihan model sebelum pembaruan bobot dilakukan (Yang & Shami, 2020). *Hyperparameter* menentukan stabilitas optimisasi, arah pembelajaran, serta kemampuan generalisasi model terhadap data baru. Optimisasi dilakukan melalui evaluasi sistematis terhadap berbagai kombinasi nilai untuk memperoleh konfigurasi terbaik, dengan metode seperti *grid search*, *random search*, dan pendekatan berbasis probabilistik (Ali dkk., 2025). Keberhasilan *tuning* sangat bergantung pada ketelitian evaluasi dan pemahaman fungsi tiap *hyperparameter*.

Berbagai *hyperparameter* memiliki peran berbeda dalam menentukan dinamika pelatihan model. Parameter yang umum disesuaikan antara lain sebagai berikut (Julianto dkk., 2022).

1. *Batch size* (ukuran *batch*) adalah jumlah sampel yang diproses dalam satu tahap pembaruan bobot. Nilai besar memberikan gradien lebih stabil dan mempercepat perhitungan, namun membutuhkan memori tinggi. Nilai kecil meningkatkan variasi gradien dan dinamika pembelajaran

2. Jumlah *epoch* menunjukkan frekuensi seluruh data pelatihan digunakan untuk memperbarui bobot. Nilai *epoch* yang rendah menyebabkan model belum konvergen, sedangkan nilai yang terlalu tinggi meningkatkan risiko *overfitting*.

3. *Learning rate* mengatur besar langkah pembaruan bobot pada setiap iterasi optimisasi. Nilai tinggi membuat pembelajaran tidak stabil, sedangkan nilai rendah memperlambat konvergensi.

4. *Dropout* merupakan teknik regularisasi dengan menonaktifkan sebagian unit jaringan selama pelatihan untuk mengurangi *overfitting* berkurang. Nilai *dropout* yang terlalu tinggi menurunkan kapasitas pembelajaran, sedangkan nilai yang terlalu rendah mengurangi efektivitas regularisasi.

5. *Hidden dimension* adalah jumlah unit tersembunyi dalam lapisan representasi yang menentukan kapasitas model dalam mempelajari pola kontekstual. Nilai *hidden dimension* yang besar memperkaya representasi, namun meningkatkan kompleksitas komputasi. Nilai yang terlalu kecil membatasi kemampuan model dalam mengenali hubungan antarfitur.

6. Jumlah layer LSTM menentukan kedalaman jaringan dalam mempelajari pola hierarkis data sekuensial. Penambahan layer memperluas kapasitas representasi namun menambah waktu pelatihan dan risiko ketidakstabilan gradien.

7. *Dense dimension* merupakan jumlah unit pada lapisan padat yang meneruskan representasi ke keluaran. Nilai besar memperluas ruang transformasi fitur namun menambah beban komputasi sedangkan nilai kecil membatasi akurasi pemetaan.

Hyperparameter tuning menempati posisi strategis dalam pengembangan model pembelajaran mendalam karena menentukan kualitas optimisasi dan kemampuan generalisasi. Pengaturan yang tepat meningkatkan efisiensi pelatihan serta

menghasilkan performa prediktif yang stabil, menjadikannya fondasi penting agar model bekerja sesuai karakteristik data dan tujuan pemodelan.

2.10.1 *Adaptive Moment Estimation* (Adam) *Optimizer*

Optimizer berfungsi memperbarui parameter model dengan meminimalkan fungsi *cost* yang merepresentasikan kualitas prediksi. Pemilihan *optimizer* berperan penting dalam menentukan akurasi model karena proses optimisasi mengatur pembaruan bobot dan laju pembelajaran secara sistematis sepanjang pelatihan. Salah satu algoritma yang paling banyak digunakan adalah Adam, yang mengintegrasikan estimasi adaptif berdasarkan informasi gradien dan iterasi sebelumnya (Reyad dkk., 2023). Mekanisme ini memungkinkan laju pembelajaran menyesuaikan diri secara dinamis, sehingga proses pembaruan parameter berlangsung stabil dan efisien.

Adam menggabungkan estimasi momen pertama dan kedua untuk menentukan arah pembaruan parameter secara adaptif terhadap variasi gradien. Pendekatan ini menyesuaikan skala pembaruan pada setiap parameter secara proporsional terhadap perubahan gradien sekaligus menjaga stabilitas melalui koreksi bias. Efisiensi meningkat karena metode ini bersifat stokastik dengan kebutuhan memori rendah dan hanya memanfaatkan gradien orde pertama. Kombinasi adaptivitas, stabilitas, dan efisiensi menjadikan Adam sangat efektif untuk berbagai arsitektur *deep learning* berdimensi tinggi yang memerlukan konvergensi cepat serta performa prediktif optimal.

2.10.2 *Early Stopping*

Menurut Anam dkk. (2024), *early stopping* merupakan teknik regularisasi yang mengendalikan proses pelatihan jaringan saraf dengan menghentikan pelatihan secara terukur sebelum mencapai batas *epoch* maksimum. Pemantauan performa dilakukan secara konsisten pada data validasi untuk menilai keberlanjutan peningkatan pembelajaran. Penurunan performa ditandai oleh kenaikan *loss* atau penurunan akurasi validasi, yang mengindikasikan mulai terbentuknya pola tidak relevan pada data latih. Parameter *patience* digunakan untuk menentukan toleransi iterasi sebelum penghentian dilakukan, guna mencegah model kehilangan kemampuan generalisasi.

Penerapan *early stopping* menghasilkan model yang stabil dan efisien karena pelatihan dihentikan saat peningkatan performa tidak lagi signifikan. Efisiensi komputasi tetap terjaga dengan mencegah eksekusi *epoch* tambahan yang tidak meningkatkan kualitas prediksi. Keputusan penghentian yang selaras dengan dinamika performa validasi memungkinkan model mencapai titik optimum pelatihan tanpa risiko *overfitting*. Dengan demikian, transisi pembelajaran menuju kondisi akhir berlangsung terkontrol, dan model tetap mempertahankan kemampuan adaptif terhadap data yang tidak digunakan dalam pelatihan.

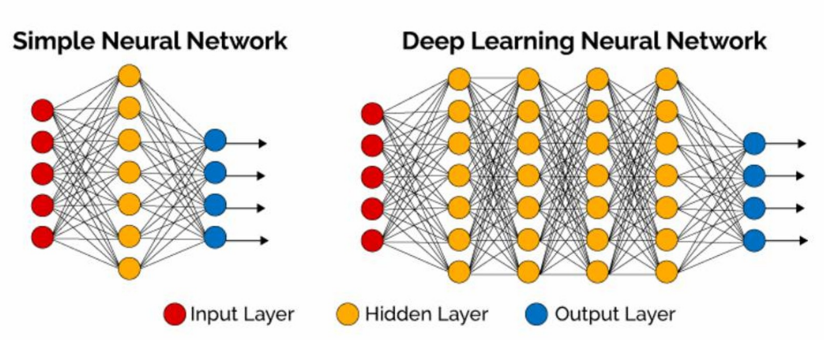
2.11 *Machine Learning*

Machine learning merupakan cabang kecerdasan buatan yang berfokus pada pengembangan algoritma yang mampu mempelajari pola data tanpa pemrograman eksplisit (Soori dkk., 2023). Sistem memperoleh kemampuan inferensial melalui proses pembelajaran berbasis data sehingga dapat meningkatkan kinerja secara bertahap. Pendekatan ini memanfaatkan metode statistik dan komputasional untuk membangun model yang mampu menggeneralisasi informasi dari data pelatihan.

Terdapat tiga kategori utama pembelajaran dalam *machine learning* yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. *Supervised learning* menggunakan data berlabel untuk membimbing model dalam menghasilkan prediksi melalui proses klasifikasi maupun regresi. Paradigma ini menempatkan label sebagai acuan untuk memetakan hubungan antara *input* dan *output* sehingga model mampu mengenali pola secara terarah. *Unsupervised learning* memproses data tanpa label melalui identifikasi struktur laten yang muncul secara alami, termasuk pengelompokan dan reduksi dimensi, sehingga pola dapat ditemukan tanpa pengawasan eksternal. *Reinforcement learning* mengembangkan strategi pengambilan keputusan berdasarkan interaksi model dengan lingkungan melalui sinyal penguatan yang mendorong pemilihan tindakan yang memaksimalkan nilai penghargaan (Akpinar, 2023). Ketiga pendekatan tersebut digunakan dalam berbagai konteks seperti pengenalan teks, pengenalan wajah, dan prediksi nilai. Tugas *Named Entity Recognition* (NER) termasuk dalam ranah *supervised learning* karena proses pelatihannya memerlukan data yang telah dianotasi dengan label.

2.12 *Deep Learning*

Deep learning merupakan cabang lanjutan dari *machine learning* yang terinspirasi dari cara kerja otak manusia dalam memproses informasi. Teknologi ini memungkinkan sistem mempelajari representasi data secara hierarkis melalui jaringan saraf tiruan berlapis banyak, di mana setiap lapisan mengekstraksi fitur dari data mentah hingga membentuk abstraksi tingkat tinggi yang mampu mengenali pola kompleks, hubungan nonlinier, dan karakteristik tersembunyi. Mekanisme pembelajaran dilakukan melalui penyesuaian bobot secara iteratif menggunakan propagasi balik (*backpropagation*), menghasilkan representasi stabil dan kemampuan generalisasi terhadap data baru (Gajiwala, 2025). Gambar 3 menunjukkan struktur umum jaringan *deep learning*.



Gambar 3. Arsitektur jaringan saraf sederhana dan jaringan *deep learning* (Nurhakiki & Yahfizham, 2024).

Jaringan *deep learning* berbeda dari jaringan saraf sederhana karena memiliki lebih banyak lapisan tersembunyi, memungkinkan pembelajaran hierarkis dan representasi fitur yang lebih informatif untuk klasifikasi atau prediksi. Pendekatan ini otomatis mengekstraksi fitur relevan tanpa rekayasa manual, sehingga adaptif terhadap variasi data dan efektif untuk permasalahan kompleks maupun data tidak terstruktur seperti citra, suara, dan teks.

Beberapa arsitektur *deep learning* yang umum digunakan antara lain *Convolutional Neural Network* (CNN), *Recurrent Neural Network* (RNN), *Long Short-Term Memory* (LSTM), dan Transformer. CNN unggul untuk data citra 2D, RNN dan LSTM efektif pada data sekuensial, sedangkan Transformer mampu menangkap konteks dan ketergantungan jarak jauh dalam teks (Shiri dkk., 2023). Setiap neuron

pada lapisan memproses *input* melalui bobot dan fungsi aktivasi, menentukan pengaruh sinyal terhadap lapisan berikutnya. Model dengan lebih dari tiga lapisan mampu merepresentasikan fitur secara berlapis dan otomatis, sehingga penerapannya pada pengenalan pola visual, pengenalan suara, pemrosesan bahasa alami, dan deteksi objek menunjukkan keunggulan *deep learning* dibandingkan *machine learning* konvensional.

2.12.1 Fungsi Aktivasi *Rectified Linear Unit* (ReLU)

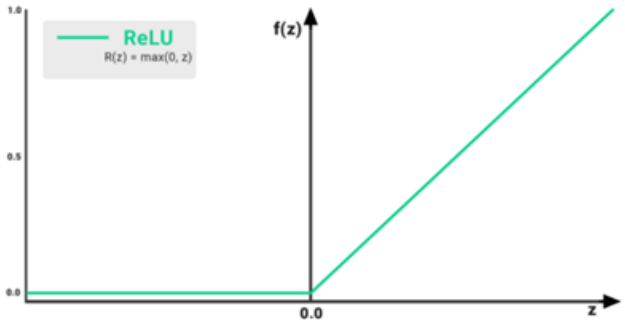
Fungsi aktivasi merupakan komponen krusial dalam jaringan saraf tiruan yang mengubah sinyal *input* menjadi *output* agar informasi dari satu lapisan dapat diteruskan ke lapisan berikutnya secara terstruktur (Kurniawan dkk., 2024). Fungsi ini memperkenalkan non-linearitas sehingga jaringan saraf mampu mengenali pola kompleks yang tidak dapat ditangkap oleh transformasi linear sederhana. Beberapa fungsi aktivasi umum meliputi sigmoid, softmax, tangent hiperbolik (tanh), *Rectified Linear Unit* (ReLU), dan *Gaussian Error Linear Unit* (GELU), masing-masing dengan karakteristik tersendiri dalam mendukung performa jaringan.

ReLU merupakan fungsi aktivasi non-linear yang paling banyak digunakan karena kesederhanaan dan efisiensinya (Gustineli, 2022). Fungsi ini mengubah nilai *input* negatif menjadi nol dan mempertahankan nilai positif sebagaimana didefinisikan pada Persamaan 1 sebagai berikut:

$$f(x) = \max(0, x) = \begin{cases} x, & \text{jika } x \geq 0, \\ 0, & \text{jika } x < 0 \end{cases} \tag{1}$$

Keterangan:

x = nilai *input* ke neuron.
$f(x)$ = *output* neuron.



Gambar 4. Fungsi Aktivasi ReLU (Purwitasari & Soleh, 2022).

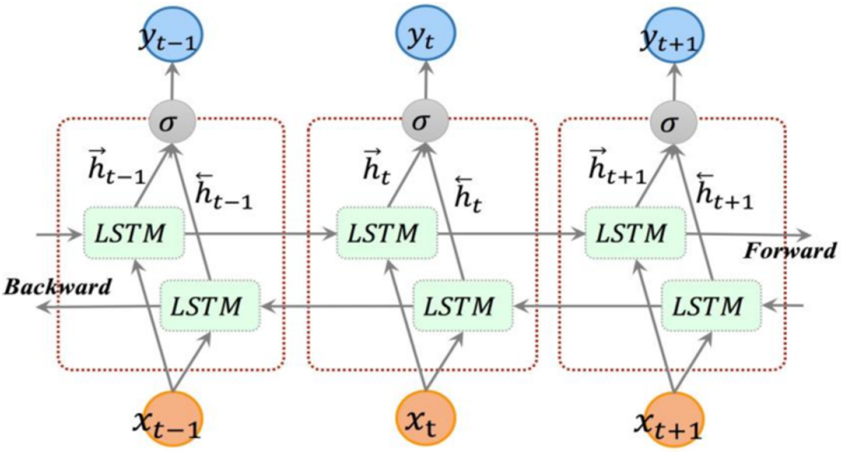
Fungsi aktivasi ReLU diilustrasikan oleh Gambar 4. Sifat non-linear yang diperkenalkan ReLU memungkinkan model mempelajari pola kompleks sekaligus mendukung *sparsity output*, yaitu kondisi di mana sebagian *neuron* tidak aktif pada satu waktu sehingga meningkatkan efisiensi dan kemampuan generalisasi model. Keunggulan ReLU terlihat pada percepatan konvergensi selama pelatihan karena gradien tetap positif untuk *input* bernilai positif, mengurangi risiko *vanishing gradient* yang umum terjadi pada fungsi sigmoid dan tanh. Variasi ReLU, seperti *Leaky* ReLU dan *Parametric* ReLU, mengatasi masalah *neuron* mati yang tidak memperbarui bobot selama pelatihan. Implementasi ReLU efektif dalam berbagai aplikasi, termasuk pengenalan citra, pemrosesan bahasa alami, dan prediksi deret waktu, sehingga menjadikannya pilihan utama dalam desain arsitektur *deep learning* modern.

Jaringan saraf tiruan meniru struktur dan fungsi neuron di otak manusia, di mana informasi diproses dan diteruskan antar lapisan hingga mencapai *output* akhir. Fungsi aktivasi, khususnya ReLU, memungkinkan lapisan tersembunyi menangani hubungan non-linear antara *input* dan *output* secara efektif. Fungsi ini juga berperan dalam perhitungan bobot dan bias, pembaruan gradien, serta mendukung berbagai aplikasi jaringan saraf, termasuk klasifikasi objek, pengenalan suara, prediksi cuaca, deteksi penyakit, dan penerjemahan mesin.

2.12.2 *Bidirectional Long Short-Term Memory* (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) merupakan pengembangan dari arsitektur *Long Short-Term Memory* (LSTM) yang dirancang untuk mengatasi

keterbatasan memori jangka panjang pada jaringan saraf berulang (RNN) konvensional. BiLSTM memproses data secara dua arah, yaitu dari *input* awal ke akhir (*forward*) dan dari *input* akhir ke awal (*backward*), sehingga setiap *timestep* memiliki konteks informasi dari kedua arah (Aung dkk., 2023). Arsitektur ini memungkinkan model memahami dependensi jangka panjang dan hubungan kontekstual antar elemen data dengan lebih akurat, khususnya pada data berurutan seperti teks, suara, dan sinyal waktu. Arsitektur BiLSTM ditunjukkan pada Gambar 5.



Gambar 5. Arsitektur Model BiLSTM (Wiujianna dkk., 2025).

Setiap unit LSTM dalam BiLSTM terdiri dari *input gate*, *output gate*, dan *forget gate* yang mengontrol aliran informasi serta mempertahankan memori penting selama pelatihan. Mekanisme ini mencegah hilangnya informasi relevan dan meminimalkan efek *vanishing gradient* yang umum terjadi pada RNN standar. Secara matematis, *output forward* dan *backward* pada *timestep* t dapat dituliskan sebagai berikut.

$$\overrightarrow{h}_t = \mathrm{LSTM}(x_t, \overrightarrow{h}_{t-1}) \tag{2}$$

$$\overleftarrow{h}_t = \mathrm{LSTM}(x_t, \overleftarrow{h}_{t+1}) \tag{3}$$

$$h_t = [\overrightarrow{h}_t ; \overleftarrow{h}_t] \tag{4}$$

Persamaan 2 menunjukkan proses *forward* LSTM, Persamaan 3 menunjukkan proses *backward* LSTM, dan Persamaan 4 merupakan *output* akhir BiLSTM yang merupakan gabungan dari kedua arah.

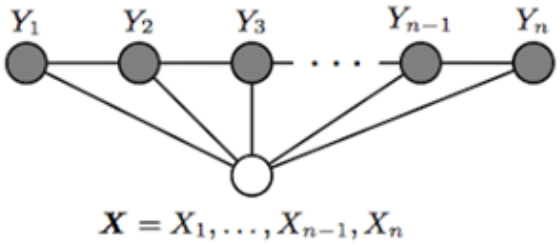
Keterangan variabel:

$h_t =$ *output gate* LSTM dua arah
$\overrightarrow{h_t} =$ *output* LSTM maju (*forward*)
$\overleftarrow{h_t} =$ *output* LSTM mundur (*backward*)
$x_t =$ *input* pada waktu ke-t
$\overrightarrow{h_{t-1}} =$ *hidden state* pada *timestep* sebelumnya
$\overleftarrow{h_{t+1}} =$ *hidden state* pada *timestep* berikutnya

BiLSTM menggabungkan *output* dari kedua arah pemrosesan, menghasilkan representasi fitur yang lebih kaya dan komprehensif dibandingkan LSTM *unidirectional*, sehingga meningkatkan performa model dalam prediksi dan klasifikasi data sekuensial. Implementasi BiLSTM terbukti efektif dalam berbagai aplikasi, termasuk pemrosesan bahasa alami, analisis sentimen, pengenalan ucapan, dan prediksi deret waktu. Model ini mampu menangkap konteks sebelum dan sesudah suatu *token* dalam teks, meningkatkan akurasi dan kemampuan generalisasi, serta memungkinkan jaringan saraf menghasilkan representasi data yang lebih informatif untuk mendukung pengambilan keputusan pada berbagai tugas berbasis sekuensial.

2.12.3 *Conditional Random Field* (CRF)

Conditional Random Field (CRF) merupakan model probabilistik diskriminatif yang digunakan untuk memprediksi data berurutan dengan mempertimbangkan ketergantungan antar label. Model ini memanfaatkan informasi kontekstual dari label sebelumnya untuk meningkatkan akurasi prediksi, berbeda dengan model klasifikasi tradisional yang memperlakukan setiap sampel secara independen (Patil dkk., 2020). CRF memodelkan hubungan antar label melalui representasi graf, sehingga memungkinkan prediksi urutan yang lebih konsisten dan akurat pada data sekuensial. Pendekatan ini telah terbukti efektif pada berbagai tugas pemrosesan bahasa alami (NLP), termasuk *part-of-speech tagging*, pengenalan entitas bernama (NER), ekstraksi kutipan, dan identifikasi sumber opini. Arsitektur CRF ditunjukkan pada Gambar 6.



Gambar 6. Arsitektur CRF (Ketmaneechairat & Maliyaem, 2020).

Pada *linear chain* CRF, probabilitas kondisional dari urutan label y terhadap urutan *input* x dirumuskan pada Persamaan 5 berikut:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right) \tag{5}$$

Persamaan ini memiliki keterangan sebagai berikut:

$p(y \mid x)$	= probabilitas urutan label y diberikan urutan *input* x
$Z(x)$	= fungsi normalisasi yang menjamin total probabilitas sama dengan 1
λ_k	= bobot untuk fitur ke-k, menunjukkan kontribusi fitur terhadap prediksi akhir
$f_k(y_t, y_{t-1}, x_t)$	= fungsi fitur yang mengukur hubungan antara label saat ini, label sebelumnya, dan *input* pada *timestep* t

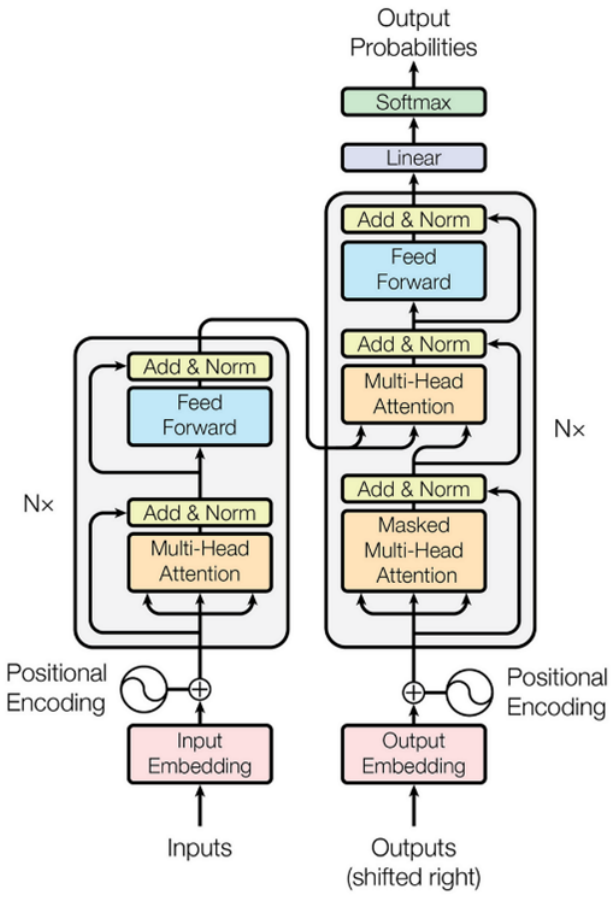
Model *linear chain* CRF mampu menangkap ketergantungan antar label dan memanfaatkan berbagai fitur dari urutan observasi, termasuk informasi kontekstual kata atau fitur statistik tambahan dari korpus pelatihan. Pendekatan ini memungkinkan sistem menghasilkan prediksi yang lebih stabil dan konsisten, serta memperbaiki hasil yang tidak sesuai melalui proses pasca-pemrosesan. Integrasi CRF dengan model pembelajaran mendalam, seperti BiLSTM-CRF, meningkatkan performa sistem NER dan pemrosesan data sekuensial yang kompleks dengan memanfaatkan representasi fitur yang lebih informatif.

2.13 *Transformer*

Transformer merupakan arsitektur *deep learning* yang menjadi tonggak penting dalam pemrosesan bahasa alami. Arsitektur ini menggantikan ketergantungan model

RNN dan CNN terhadap pemrosesan berurutan melalui mekanisme *self-attention* yang sepenuhnya paralel. Model ini mampu menganalisis seluruh *token* dalam urutan secara simultan sehingga mengatasi keterbatasan paralelisasi pada RNN yang memerlukan propagasi langkah demi langkah. Pendekatan tersebut meningkatkan efisiensi pelatihan, kemampuan menangkap dependensi jarak jauh, serta stabilitas representasi data (Asri & Kuswardani, 2024).

Mekanisme *self-attention* menjadi inti arsitektur *Transformer* karena merepresentasikan hubungan antar *token* dengan memberikan bobot lebih besar pada *token* yang relevan dalam konteks. Mekanisme ini menghubungkan posisi kata dalam kalimat secara global, menghasilkan representasi konteks yang lebih informatif. Efektivitas *self-attention* menjadikan *Transformer* unggul dibandingkan pendekatan berbasis rekuren maupun konvolusi. Arsitektur *Transformer* ditunjukkan pada Gambar 7.



Gambar 7. Arsitektur Transformer (Vaswani dkk., 2017).

Transformer memiliki dua komponen utama, yaitu *encoder* dan *decoder*. *Encoder* memproses urutan *input* $X = (x_1, \ldots, x_N)$ dan menghasilkan representasi laten $Z = (z_1, \ldots, z_N)$. *Decoder* kemudian menghasilkan urutan keluaran $Y = (y_1, \ldots, y_M)$ secara *autoregressive* dengan memanfaatkan representasi Z dan keluaran sebelumnya $Y_{(M-1)} = (y_1, \ldots, y_{M-1})$. Struktur ini memungkinkan model menghasilkan keluaran yang konsisten secara kontekstual pada setiap langkah prediksi.

Encoder terdiri atas enam lapisan identik, masing-masing memiliki dua sublapisan, yaitu *multi-head self-attention* dan *feed-forward network*. *Decoder* juga memiliki enam lapisan identik, dengan tambahan satu sublapisan *cross-attention* yang menghubungkan *decoder* dengan representasi *encoder*. Mekanisme *masked self-attention* diterapkan dalam *decoder* untuk menjaga proses *autoregressive* sehingga model tidak mengakses *token* masa depan. Setiap sublapisan dilengkapi dengan *residual connection* dan *layer normalization* untuk menjaga stabilitas aliran informasi.

Mekanisme*multi-head attention* memungkinkan model menangkap variasi hubungan antar *token* melalui beberapa *head* perhatian yang bekerja secara paralel. Setiap *head* memetakan *input* menjadi tiga komponen, yaitu *query*, *key*, dan *value*. Ketiganya digunakan untuk menghitung skor perhatian yang menentukan *token* relevan dalam konteks. *Multi-head attention* memperkaya representasi karena setiap *head* menyoroti pola dependensi yang berbeda.

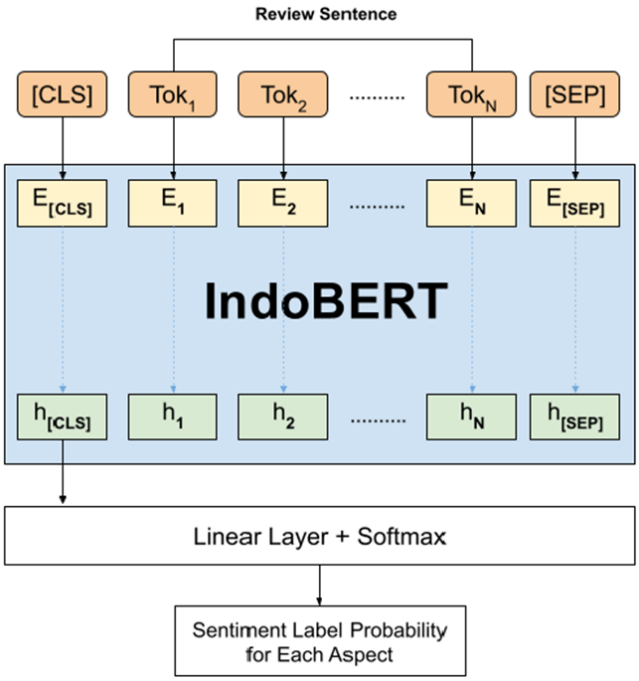
Transformer mengadopsi *positional encoding* untuk memberikan informasi posisi *token* dalam urutan, karena arsitektur ini tidak memiliki sifat berurutan seperti RNN. *Positional encoding* mempertahankan informasi struktur kalimat sehingga proses perhatian tetap mempertimbangkan urutan *token*. Integrasi antara *self-attention*, *feed-forward network*, *residual connection*, dan *positional encoding* menjadikan *transformer* mampu memproses data teks secara efektif dan efisien.

2.13.1 *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT)

IndoBERT merupakan model pralatih berbasis arsitektur *transformer* yang dirancang khusus untuk bahasa Indonesia. Model ini dilatih menggunakan dua strategi utama,

yaitu *masked language modeling* dan *next sentence prediction*, sehingga mampu membangun pemahaman sintaktis dan semantis yang sesuai dengan karakteristik bahasa Indonesia (Koto dkk., 2020).

Proses pralatih memanfaatkan korpus berukuran besar yang mencakup berbagai domain, seperti berita daring, media sosial, artikel umum, Wikipedia, subtitel, serta data paralel. Jumlah data yang digunakan mencapai 4 miliar kata, menghasilkan representasi konteks yang stabil dan adaptif terhadap variasi bahasa formal maupun nonformal (Wilie dkk., 2020). Arsitektur IndoBERT ditunjukkan pada Gambar 8.



Gambar 8. Arsitektur Model IndoBERT (Yulianti & Nissa, 2024).

IndoBERT mengadopsi struktur *encoder transformer* sebagaimana yang digunakan pada BERT. Model ini tersedia dalam dua varian utama, yaitu IndoBERT *Base* dan IndoBERT *Large*, yang dibedakan berdasarkan jumlah lapisan *encoder*, jumlah *attention head*, serta jumlah parameter. IndoBERT Base terdiri atas 12 lapisan *encoder*, 12 *attention head*, dan lebih dari 100 juta parameter. Sementara itu, IndoBERT Large mencakup 24 lapisan *encoder*, 16 *attention head*, dan lebih dari 300 juta parameter (Nabiilah dkk., 2023). Varian yang lebih besar memiliki kapasitas representasi lebih tinggi, meskipun membutuhkan sumber daya komputasi lebih besar selama pelatihan maupun inferensi.

IndoBERT menggunakan *vocabulary* khusus bahasa Indonesia yang dikembangkan melalui pendekatan *subword tokenization*. Pendekatan ini memungkinkan model mengenali variasi kata majemuk, bentuk turunan, serta kata yang jarang muncul. Proses tersebut meningkatkan ketahanan model terhadap keragaman bentuk leksikal yang menjadi ciri khas bahasa Indonesia. Representasi berbasis *subword* juga mengatasi keterbatasan model berbasis kata penuh yang sering gagal mengenali kata baru atau kata bermorfologi kompleks.

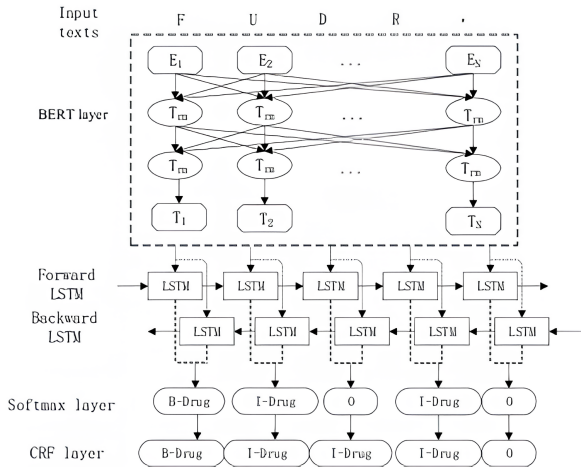
2.13.2 *Hybrid* IndoBERT-BiLSTM-CRF

Model *hybrid* IndoBERT–BiLSTM–CRF memanfaatkan representasi kontekstual IndoBERT untuk menghasilkan *embedding* kata yang sesuai dengan karakteristik bahasa Indonesia. Representasi ini bersifat lebih adaptif dibandingkan *embedding* statis karena makna setiap *token* dipahami berdasarkan konteks kemunculannya dalam kalimat. Kualitas representasi tersebut menjadi fondasi utama agar proses identifikasi entitas berlangsung lebih akurat (Nabiilah dkk., 2024).

Embedding yang dihasilkan IndoBERT diteruskan ke lapisan BiLSTM untuk menangkap pola sekuensial dari dua arah. Pemrosesan bidireksional memungkinkan model mengenali ketergantungan antar-*token* secara menyeluruh, sehingga hubungan struktural dalam kalimat dapat terpelajari dengan lebih baik. Informasi urutan ini memperkaya konteks yang telah disediakan oleh IndoBERT serta mendukung pembentukan fitur yang lebih informatif bagi tahap prediksi label.

Keluaran BiLSTM kemudian diproses oleh CRF untuk memastikan urutan label memiliki konsistensi struktural sesuai skema NER. CRF mengevaluasi hubungan antar-label sehingga kesalahan prediksi, seperti kemunculan label lanjutan tanpa label awal, dapat diminimalkan. Mekanisme ini meningkatkan ketepatan identifikasi entitas melalui kontrol global terhadap keseluruhan urutan.

Integrasi IndoBERT, BiLSTM, dan CRF menghasilkan arsitektur yang mampu menggabungkan pemahaman konteks mendalam, pemodelan urutan, serta konsistensi pelabelan. Ketiga komponen tersebut bekerja secara komplementer sehingga menghasilkan performa yang stabil dan unggul pada tugas NER bahasa Indonesia. Arsitektur penggunaan BERT–BiLSTM–CRF ditampilkan pada Gambar 9.



Gambar 9. Arsitektur Model IndoBERT-BiLSTM-CRF (Dave & Chowanda, 2024).

2.14 Evaluasi Model

Evaluasi model digunakan untuk menilai kualitas prediksi melalui perbandingan antara label aktual dan label hasil prediksi. *Confusion matrix* menjadi dasar utama penilaian karena menyajikan distribusi prediksi benar dan salah pada setiap kelas, sehingga karakter kesalahan model dapat dipetakan secara sistematis. Informasi ini memberikan gambaran menyeluruh mengenai kemampuan model dalam mengenali pola data dan membedakan setiap kelas secara konsisten.

2.14.1 *Confusion Matrix Biner*

Confusion matrix biner digunakan pada sistem klasifikasi dua kelas dengan struktur dua baris dan dua kolom, masing-masing mewakili label aktual dan label prediksi. Setiap sel matriks menunjukkan hubungan antara kondisi aktual dan hasil prediksi sehingga karakteristik kesalahan dapat dianalisis secara terarah (Sathyanarayanan & Tantri, 2024). Pendekatan ini menyediakan landasan evaluasi yang kuat terhadap performa model dalam membedakan kelas positif dan negatif. *Confusion matrix biner* ditunjukkan pada Tabel 3.

Keterangan:
TP = Data aktual positif yang diprediksi positif.
TN = Data aktual negatif yang diprediksi negatif.

Tabel 3. *Confusion Matrix Biner*.

Actual Class	***Prediction Class***	
	Positive	***Negative***
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

FP = Data aktual negatif yang diprediksi positif.
FN = Data aktual positif yang diprediksi negatif.

Analisis biner memungkinkan perhitungan metrik evaluasi melalui hubungan langsung antara empat komponen tersebut (Perdana & Adikara, 2025).

1. *Accuracy* menunjukkan proporsi prediksi yang sesuai dengan kondisi aktual terhadap seluruh observasi (Vakili dkk., 2020). Metrik ini menilai frekuensi model menghasilkan prediksi benar, baik untuk kelas positif maupun negatif. Nilai *accuracy* semakin tinggi apabila model mampu meminimalkan kesalahan FP dan FN, serta memaksimalkan prediksi benar pada TP dan TN. Rumus *accuracy* dirumuskan sebagai:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. *Precision* menunjukkan tingkat ketepatan prediksi positif terhadap seluruh prediksi yang dinyatakan positif oleh model (Vakili dkk., 2020). Metrik ini menggambarkan seberapa konsisten model menghindari kesalahan FP. Nilai *precision* tinggi mengindikasikan bahwa sebagian besar prediksi positif merupakan prediksi yang benar. Rumus *precision* dihitung sebagai:

$$Precision = \frac{TP}{TP + FP}$$

3. *Recall* menggambarkan kemampuan model dalam mengenali seluruh data positif aktual (Vakili dkk., 2020). Metrik ini menunjukkan seberapa efektif model menghindari kesalahan FN. Nilai *recall* tinggi mencerminkan kemampuan deteksi positif yang kuat, terutama pada konteks yang menuntut minimisasi kehilangan data positif. Rumus *recall* dihitung sebagai:

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

4. F1-*score* merupakan rata-rata harmonik dari *precision* dan *recall* yang memberikan evaluasi menyeluruh terhadap keseimbangan antara ketepatan dan kelengkapan prediksi (Vakili dkk., 2020). Metrik ini menjadi indikator representatif ketika *precision* dan *recall* menunjukkan nilai yang tidak seimbang, karena F1-*score* menurunkan nilai keseluruhan ketika salah satu komponen berada pada tingkat rendah. Rumus F1-*score* dinyatakan:

$$F1\text{-}score = \frac{2 \times Precision \times \text{Recall}}{Precision + Recall}$$

2.14.2 *Confusion Matrix Multiclass*

Confusion matrix multiclass digunakan pada sistem klasifikasi yang melibatkan lebih dari dua kelas. Matriks ini berbentuk N×N, dengan baris menunjukkan label aktual dan kolom menunjukkan label prediksi. Sel diagonal utama menggambarkan jumlah prediksi benar pada setiap kelas, sedangkan sel di luar diagonal menunjukkan kesalahan model dalam membedakan satu kelas dari kelas lainnya. Struktur ini memberikan pemetaan kinerja yang lebih mendalam karena setiap kelas dianalisis berdasarkan distribusi prediksinya (Siregar dkk., 2023). Variasi penyajian diperlukan untuk meningkatkan ketelitian interpretasi terutama ketika jumlah data pada setiap kelas tidak seimbang.

Normalized confusion matrix menampilkan proporsi prediksi pada setiap kelas melalui normalisasi nilai pada baris atau kolom sehingga menghasilkan rentang antara 0 hingga 1. Representasi ini menunjukkan persentase relatif terhadap total data pada masing-masing kelas, sehingga interpretasi tetap stabil meskipun distribusi data tidak seimbang. Pendekatan ini memperjelas kecenderungan model dalam mengklasifikasikan kelas tertentu serta menegaskan pola kesalahan yang tidak tampak pada matriks tanpa normalisasi. Visualisasi terstandarisasi tersebut meningkatkan akurasi interpretasi distribusi prediksi dan memperkuat analisis performa klasifikasi *multiclass*.

Evaluasi *multiclass* menggunakan nilai *True Positive* (TP), *False Positive* (FP), dan *False Negative* (FN) untuk setiap kelas. Nilai *True Negative* (TN) tidak dihitung karena tidak memiliki peran dalam penilaian per kelas. Pendekatan ini memungkinkan pengukuran performa secara agregatif melalui rata-rata mikro, makro,

atau berbobot, sehingga kontribusi setiap kelas tetap proporsional. *Confusion matrix multiclass* ditunjukkan pada Tabel 4.

Tabel 4. *Confusion Matrix Multiclass*.

Actual Classes	***Predicted Classification***			
	a	**b**	**c**	**d**
A	TN	FP	TN	TN
B	FN	TP	FN	FN
C	TN	FP	TN	TN
D	TN	FP	TN	TN

Menurut Grandini dkk. (2020), metrik evaluasi dalam konteks *multiclass* tetap menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Perhitungan dilakukan melalui agregasi nilai TP, FP, dan FN dari seluruh kelas agar evaluasi tetap konsisten dengan formulasi biner.

Rumus *accuracy multiclass* dinyatakan sebagai:

$$Accuracy = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FP_i + FN_i)}$$

Sementara itu, *precision*, *recall*, dan *F1-score* per kelas dihitung menggunakan:

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

dengan i menyatakan indeks kelas ke-i.

2.14.3 *Overfitting* dan *Underfitting*

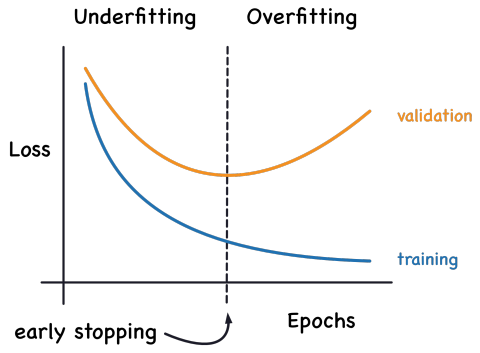
Overfitting dan *underfitting* merupakan kondisi yang mencerminkan kualitas generalisasi model terhadap data di luar data pelatihan. Generalisasi mengacu pada kemampuan model dalam mengenali pola yang relevan sehingga dapat diterapkan pada data baru. Evaluasi kondisi ini dilakukan melalui analisis *learning curve*,

khususnya hubungan antara *training loss* dan *validation loss* sepanjang proses pelatihan. Pola perubahan kedua kurva menjadi dasar objektif dalam menilai kemampuan model secara sistematis.

Overfitting secara umum didefinisikan sebagai kondisi ketika model terlalu menyesuaikan diri terhadap data pelatihan hingga menangkap detail yang tidak relevan, termasuk *noise*. Kondisi ini menyebabkan kinerja model sangat baik pada data pelatihan, namun menurun pada data baru. Dalam *learning curve*, *overfitting* ditandai oleh penurunan *training loss* secara terus-menerus, sementara *validation loss* meningkat setelah mencapai titik minimum. Pola tersebut menunjukkan adanya kesenjangan antara kinerja pada data pelatihan dan data validasi yang mencerminkan penurunan kemampuan generalisasi (Li dkk., 2024).

Underfitting secara umum didefinisikan sebagai kondisi ketika model belum mampu menangkap pola dasar dalam data. Kondisi ini terjadi ketika kompleksitas model tidak memadai atau proses pelatihan belum optimal. Model pada kondisi ini menunjukkan kinerja rendah baik pada data pelatihan maupun data validasi. Dalam *learning curve*, *underfitting* ditunjukkan oleh nilai *training loss* dan *validation loss* yang sama-sama tinggi serta tidak mengalami penurunan signifikan (Liu dkk., 2023).

Pola hubungan antara *training loss* dan *validation loss* dapat divisualisasikan melalui grafik *learning curve* sebagaimana ditunjukkan pada Gambar 10. Visualisasi ini memperlihatkan perubahan nilai *loss* terhadap *epoch* serta perbedaan karakteristik pada setiap fase pelatihan.



Gambar 10. *Overfitting* dan *underfitting* pada kurva *loss* (Montesinos-López dkk., 2022).

Berdasarkan Gambar 10, pada tahap awal pelatihan kedua kurva berada pada nilai tinggi yang menunjukkan kondisi *underfitting*. Seiring bertambahnya *epoch*, *training loss* dan *validation loss* menurun secara bersamaan yang menunjukkan proses pembelajaran berlangsung dengan baik. Kondisi optimal dicapai saat *validation loss* mencapai nilai minimum dan selisih antara kedua kurva relatif kecil. Setelah titik tersebut, *validation loss* meningkat sementara *training loss* terus menurun, yang menunjukkan terjadinya *overfitting*.

Rentang atau indikator kondisi model ditentukan berdasarkan hubungan nilai *training loss* dan *validation loss* pada setiap *epoch*, bukan berdasarkan tampilan visual grafik. Kondisi *underfitting* terjadi ketika kedua nilai *loss* masih tinggi. Kondisi optimal ditandai oleh nilai *validation loss* minimum dengan selisih kecil terhadap *training loss*. Kondisi *overfitting* ditunjukkan oleh peningkatan *validation loss* yang disertai penurunan *training loss*, sehingga menghasilkan kesenjangan yang semakin besar antara kedua kurva. Titik minimum *validation loss* menjadi batas penting dalam menentukan *epoch* terbaik serta digunakan sebagai acuan untuk menjaga kemampuan generalisasi model.

Dalam praktiknya, tingkat *overfitting* dapat bervariasi, mulai dari ringan hingga berat. Tingkatan ini umumnya ditentukan berdasarkan besar kecilnya kesenjangan antara *training loss* dan *validation loss*. Kesenjangan yang relatif kecil menunjukkan *overfitting* ringan, sedangkan kesenjangan yang semakin besar mengindikasikan *overfitting* yang lebih signifikan.

2.15 *Receiver Operating Characteristic – Area Under the Curve* (ROC–AUC)

Receiver Operating Characteristic (ROC) merupakan metode evaluasi yang digunakan untuk menilai kinerja model klasifikasi berdasarkan kemampuan prediksi pada berbagai nilai ambang keputusan. Kurva ROC menggambarkan kinerja klasifikasi tanpa mempertimbangkan distribusi kelas maupun jenis kesalahan yang terjadi. Pendekatan ini memungkinkan evaluasi performa model secara menyeluruh dan independen terhadap kondisi data.

Kurva ROC dibentuk melalui pemetaan *True Positive Rate* (TPR) pada sumbu vertikal dan *False Positive Rate* (FPR) pada sumbu horizontal. TPR merepresentasikan

proporsi data positif yang berhasil diklasifikasikan secara benar, sedangkan FPR menunjukkan proporsi data negatif yang keliru diklasifikasikan sebagai positif. Hubungan antara kedua nilai tersebut mencerminkan keseimbangan antara kemampuan deteksi kelas positif dan tingkat kesalahan pada kelas negatif.

Penilaian kuantitatif terhadap kurva ROC dilakukan melalui perhitungan *Area Under the Curve* (AUC), yaitu luas area di bawah kurva ROC. Nilai AUC berada pada rentang 0,0 hingga 1,0 dan digunakan sebagai indikator kualitas pemisahan kelas oleh model klasifikasi (Natzir, 2023). Nilai AUC yang semakin besar menunjukkan kemampuan pemisahan kelas yang semakin baik, sedangkan nilai AUC sebesar 0,5 menunjukkan performa yang setara dengan prediksi acak.

Nilai AUC juga digunakan untuk mengelompokkan tingkat kinerja model klasifikasi. Rentang nilai 0,9 hingga 1,0 menunjukkan kinerja sangat baik, rentang 0,8 hingga 0,9 menunjukkan kinerja baik, rentang 0,7 hingga 0,8 menunjukkan kinerja rata-rata, rentang 0,6 hingga 0,7 menunjukkan kinerja rendah, serta rentang 0,5 hingga 0,6 menunjukkan kegagalan klasifikasi (Fakih dkk., 2025). Pengelompokan ini memberikan acuan objektif dalam interpretasi hasil evaluasi model.

Keunggulan utama ROC–AUC terletak pada kemampuannya dalam merepresentasikan pengaruh variasi ambang keputusan terhadap kinerja model secara visual dan numerik. Karakteristik tersebut menjadikan ROC–AUC tidak bergantung pada satu nilai ambang tertentu dan relatif stabil pada kondisi ketidakseimbangan kelas. Oleh sebab itu, ROC–AUC banyak digunakan sebagai metrik evaluasi dalam berbagai kajian klasifikasi.

2.16 *Error Analysis*

Error analysis merupakan metode yang digunakan untuk menelaah kesalahan prediksi dengan meninjau hubungan antara label aktual dan label yang dihasilkan model. Pendekatan ini memetakan karakteristik kesalahan secara sistematis sehingga faktor penyebab ketidaktepatan prediksi dapat diidentifikasi secara jelas. Melalui analisis ini, diperoleh pemahaman mengenai pola data yang tidak dikenali model serta keterbatasan representasi yang muncul selama proses pelatihan.

Kesalahan prediksi pada model klasifikasi dapat berupa *misclassification*, *boundary error*, dan *confusion* antarkelas. *Misclassification* terjadi ketika model menetapkan label yang keliru akibat kemiripan fitur antar kelas. *Boundary error* muncul saat model tidak mampu menentukan batas entitas secara tepat sehingga *token* memperoleh label yang tidak sesuai konteks. Adapun *confusion* antarkelas terjadi ketika distribusi fitur antarlabel tumpang tindih dan menyebabkan pertukaran prediksi antar kelas.

Untuk mengidentifikasi pola kesalahan secara kuantitatif, digunakan kerangka evaluasi berbasis *confusion matrix* yang menggambarkan distribusi kesalahan berupa FP dan FN pada setiap kelas. Nilai FP menunjukkan kecenderungan model memberikan prediksi berlebih terhadap suatu label, sedangkan FN menggambarkan ketidakmampuan model mengenali *instance* yang sebenarnya termasuk dalam kelas tersebut. Analisis perbandingan FP dan FN membantu mengidentifikasi kelas yang rentan terhadap kesalahan serta pola prediksi yang kurang stabil.

Selain pendekatan numerik, visualisasi berbasis *Word Cloud* digunakan untuk memperkuat interpretasi *error analysis*. *Word Cloud* merupakan representasi visual yang menampilkan kata berdasarkan frekuensi kemunculannya, dengan ukuran kata mencerminkan tingkat kemunculan dalam korpus teks (Ibrahim dkk., 2021). Dalam konteks *error analysis*, visualisasi ini dibangun dari *token* yang berkontribusi terhadap kesalahan, khususnya pada FP dan FN, untuk menyoroti kata ambigu, bentuk serupa, atau konteks langka. Kombinasi *confusion matrix* dan *Word Cloud* memberikan pemahaman komprehensif terhadap sumber kesalahan model sehingga hasil analisis dapat menjadi dasar dalam penyempurnaan *preprocessing*, peningkatan variasi data pelatihan, penguatan fitur kontekstual, serta penyesuaian parameter model.

BAB III

METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian dilakukan pada waktu dan tempat sebagai berikut:

3.1.1 Tempat Penelitian

Penelitian ini dilaksanakan pada semester genap tahun akademik 2025/2026 di Pusat Riset Sains Data dan Informasi, Badan Riset dan Inovasi Nasional (BRIN) yang berlokasi di Jalan Sangkuriang, Kota Bandung, Provinsi Jawa Barat. Kegiatan penelitian kemudian dilanjutkan pada semester ganjil tahun akademik 2025/2026 di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung, yang beralamat di Jalan Prof. Dr. Ir. Soemantri Brojonegoro, Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Provinsi Lampung.

3.1.2 Waktu Penelitian

Kegiatan penelitian dimulai pada semester ganjil tahun akademik 2025/2026, tepatnya pada bulan September 2025. Pelaksanaan penelitian berlangsung melalui tiga tahapan utama yang saling berkesinambungan. Tahap pertama merupakan tahap persiapan, yang mencakup penentuan topik penelitian, pengumpulan referensi melalui studi literatur, persiapan data, eksplorasi data awal, serta penyusunan draft proposal penelitian. Tahap kedua merupakan tahap pelaksanaan yang berfokus pada proses pengolahan dan pemrograman data, meliputi kegiatan *input data*, *preprocessing*, pembagian data, penerapan POS *tagging*, pembuatan *embedding* menggunakan FastText, pembangunan model berbasis *pre-trained* Transformer, proses *fine-tuning*, serta evaluasi performa model berdasarkan metrik evaluasi yang

relevan dan perbandingan hasil antar model. Selanjutnya, tahap ketiga merupakan tahap penyusunan hasil penelitian yang berfokus pada analisis temuan empiris serta perumusan kesimpulan berdasarkan hasil pengujian model.

3.2 Data dan Alat Penelitian

Penelitian dilakukan dengan data dan alat sebagai berikut.

3.2.1 Data

Dataset yang digunakan dalam penelitian ini adalah InaCOVED (*Indonesian COVID-19 Online News and Entities Dataset*), yaitu kumpulan judul berita daring berbahasa Indonesia yang memuat informasi mengenai penyakit menular dengan fokus utama pada COVID-19. *Dataset* ini dikembangkan dan dikelola oleh Badan Riset dan Inovasi Nasional (BRIN). InaCOVED tidak bersifat *open-access* dan diperoleh melalui penyimpanan internal, meskipun bersumber dari berbagai portal berita daring nasional. Data berasal dari tujuh portal berita nasional, yaitu Kompas, Detik, Tempo, Tirto, Republika, Merdeka, dan Antara, yang merepresentasikan keberagaman sumber, gaya bahasa, serta cakupan informasi dalam pemberitaan kesehatan nasional. Untuk memberikan gambaran mengenai bentuk data yang digunakan, contoh judul berita dalam *dataset* InaCOVED beserta portal sumbernya ditampilkan pada Tabel 5.

Tabel 5. Contoh Dataset InaCOVED.

Title	**Portal**
WHO: Virus corona capai 1320 kasus di 10 negara	Antara
Covid-19 100 Kali Lebih Menular daripada SARS dan Flu Burung	Republika
Jokowi dan Trump Sepakat Jalin Kerjasama Peningkatan Alkes Tangani Covid-19	Merdeka
Eijkman: Cina dan Kanada Tawarkan Pengembangan Vaksin Covid-19	Tempo
Pemkab Tangerang Siapkan Lahan Makam Korban Corona Seluas 3000 Meter Persegi	Detik

Dataset InaCOVED disusun dalam format CoNLL yang umum digunakan pada penelitian pemrosesan bahasa alami, khususnya untuk tugas NER. Pada tahap

awal, struktur data mencakup dua atribut, yaitu `title_clean` dan `portal`. Atribut `title_clean` merepresentasikan judul berita yang telah melalui proses pembersihan teks, sedangkan atribut portal menunjukkan sumber media daring tempat judul berita dipublikasikan. Setelah anotasi dilakukan, data disesuaikan dengan skema CoNLL melalui penambahan label entitas pada tingkat *token*, sehingga memungkinkan proses pelatihan, validasi, dan pengujian model NER dilakukan secara sistematis. Jumlah data yang digunakan sebanyak 16.839 judul berita, seluruhnya dimanfaatkan tanpa pembatasan jumlah maupun penghapusan duplikat.

Pengumpulan data dilakukan pada periode Januari-Mei 2020, yaitu fase awal penyebaran COVID-19 di Indonesia ketika pemberitaan pandemi mendominasi media nasional. Informasi dalam judul berita mencakup perkembangan kasus, wilayah terdampak, kebijakan penanganan, kerja sama antarnegara, serta peran lembaga kesehatan. Kondisi temporal tersebut menjadikan *dataset* InaCOVED relevan untuk analisis informasi penyakit menular pada tahap awal pandemi.

Penelitian ini membatasi data pada judul berita. Judul memiliki karakteristik ringkas, padat informasi, dan minim konteks, sehingga meningkatkan kompleksitas proses pengenalan entitas. Berdasarkan data mentah sebelum tahap *preprocessing*, rata-rata panjang judul berita tercatat sebesar 10,22 *token* dengan median 10 *token*, panjang terpendek 3 *token*, dan terpanjang 27 *token*. Variasi tersebut menunjukkan keberagaman struktur kalimat serta kepadatan informasi pada teks pendek yang dianalisis.

Pada tahap awal, *dataset* InaCOVED belum dilengkapi anotasi entitas sehingga tidak dapat langsung digunakan untuk pelatihan model. Proses pelabelan entitas dilakukan secara manual terhadap seluruh judul berita menggunakan Label Studio. Proses anotasi melibatkan empat anotator yang bekerja berdasarkan panduan anotasi terstruktur untuk menjaga konsistensi pelabelan. Penelitian ini menggunakan empat kategori entitas dominan dalam pemberitaan penyakit menular, yaitu nama orang (PER), organisasi (ORG), lokasi (LOC), dan penyakit (DIS).

Dataset beranotasi dibagi menjadi data model dan data uji dengan proporsi 80% untuk data model dan 20% untuk data uji. Data model kemudian dibagi kembali menjadi 80% data latih dan 20% data validasi. Skema pembagian ini diterapkan agar

proses pelatihan, validasi, dan pengujian model secara terpisah sehingga evaluasi kinerja model dapat dilakukan secara objektif.

Keandalan hasil anotasi diuji melalui pengukuran reliabilitas antar anotator menggunakan koefisien Cohen's Kappa. Hasil pengujian menunjukkan nilai Cohen's Kappa sebesar 0,9567 dengan nilai *observed agreement* sebesar 0,9818. Berdasarkan kriteria interpretasi Landis dan Koch, nilai tersebut berada pada kategori *almost perfect agreement*, yang menunjukkan tingkat kesepakatan anotator yang sangat tinggi. Hasil pengujian ini menegaskan bahwa anotasi dilakukan secara konsisten dan *dataset* beranotasi memiliki kualitas memadai untuk mendukung pelatihan, validasi, serta pengujian model NER.

Keterbatasan data yang hanya mencakup judul berita serta periode awal pandemi tidak memengaruhi tujuan penelitian secara signifikan. *Dataset* InaCOVED tetap mampu merepresentasikan karakteristik utama pemberitaan penyakit menular serta mendukung evaluasi metode NER yang diterapkan dalam penelitian ini.

3.2.2 Alat Penelitian

Pada penelitian ini alat yang digunakan sebagai berikut:

3.2.2.1. Perangkat keras (*Hardware*)

Penelitian ini menggunakan laptop Acer Aspire A514-55 dengan tipe sistem 64-bit *Operating System* berbasis x64-*based processor*. Spesifikasi *hardware* yang digunakan adalah sebagai berikut:

- *Prosesor*: 12th Gen Intel(R) Core(TM) i5-1235U (12 CPU) @ 1.30 GHz
- *Memory*: SSD 512 GB
- RAM: 24 GB

3.2.2.2. Perangkat Lunak (*software*)

Sistem operasi yang digunakan dalam penelitian ini adalah Windows 11, dengan bahasa pemrograman *Python* versi 3.11.12. Proses pemrograman dan eksekusi kode dilakukan menggunakan *Google Colab* sebagai platform komputasi berbasis *cloud*. *Library Python* yang digunakan dalam penelitian ini disajikan pada Tabel 6.

Tabel 6. *Library Python*.

No.	*Library*	Versi	Penjelasan
1.	NumPy	2.0.2	*Library* komputasi numerik yang menyediakan array multidimensi dan fungsi matematis berperforma tinggi untuk operasi aljabar linier, transformasi Fourier, serta pembangkitan bilangan acak.
2.	Pandas	2.2.2	*Library* pengolahan data yang menyediakan struktur *DataFrame* untuk manipulasi, analisis, dan transformasi data secara efisien.
3.	PyTorch	2.9.0+cu126	*Library deep learning* berbasis Python yang mendukung grafik komputasi dinamis untuk pembangunan, pelatihan, dan evaluasi jaringan saraf tiruan.
4.	torchcrf	torchcrf	*Library* yang menyediakan lapisan *Conditional Random Field (CRF)* untuk meningkatkan performa model pada tugas *sequence labeling* seperti NER.
5.	Stanza	1.11.0	*Library* pemrosesan bahasa alami (NLP) yang menyediakan *pipeline* berbasis *neural network* meliputi tokenisasi, lemmatisasi, POS *tagging*, dan NER.
6.	Gensim	4.4.0	*Library* pemodelan topik dan pembentukan *embedding* kata seperti FastText dan Word2Vec.
7.	Transformers (*HuggingFace*)	4.57.1	*Library* pemanggilan model Transformer seperti IndoBERT untuk *embedding* kontekstual.
8.	Optuna	4.6.0	Framework optimasi *hyperparameter* otomatis berbasis *trial* untuk mencari konfigurasi model terbaik.

No.	*Library*	Versi	Penjelasan
9.	Scikit-Learn	1.6.1	*Library* pembelajaran mesin yang menyediakan algoritma ML, fungsi evaluasi, dan pembagian data.
10.	Seqeval	1.2.2	*Library* evaluasi untuk tugas *sequence labeling* berbasis token menggunakan metrik *precision*, *recall*, dan F1-*score*.
11.	Matplotlib	3.10.0	*Library* visualisasi data yang menyediakan fungsi pembuatan grafik, plot, dan diagram.
12.	Seaborn	0.13.2	*Library* visualisasi berbasis Matplotlib yang menghasilkan grafik statistik informatif dan estetis.
13.	WordCloud	wordcloud. wordcloud	*Library* visualisasi teks yang menampilkan frekuensi kata dalam bentuk *word cloud*.
14.	tqdm	4.67.1	*Library* penampil progres yang digunakan saat pemrosesan data, pelatihan model, atau pemuatan berkas.

3.3 Metode Penelitian

Penelitian ini berfokus pada analisis parameter terbaik dalam mendeteksi kesalahan label I-DIS pada tugas NER terhadap judul berita penyakit menular. Proses penelitian dilaksanakan berdasarkan alur penelitian sebagaimana ditampilkan pada Gambar 11. Adapun tahapan-tahapan penelitian dijelaskan sebagai berikut.

Berdasarkan alur penelitian pada Gambar 11, berikut penjelasan setiap tahapan yang dilakukan:

1. Penelitian diawali dengan penyusunan data berupa judul berita penyakit menular yang menjadi sumber utama analisis. *Dataset* dikumpulkan dan disusun sebagai korpus penelitian yang digunakan dalam proses pelabelan, pemrosesan, serta pengembangan model NER.

2. Tahap berikutnya adalah pelabelan manual terhadap empat entitas utama, yaitu *Person* (PER), *Organization* (ORG), *Location* (LOC), dan *Disease* (DIS) menggunakan Label Studio. Proses anotasi menerapkan skema BIO (*Begin–Inside–Outside*) sehingga setiap *token* dalam teks memperoleh label yang sesuai. Hasil tahap ini berupa *dataset* berlabel yang menjadi dasar proses

pemodelan.

3. Setelah pelabelan, dilakukan analisis eksploratif data (*Exploratory Data Analysis*/EDA) untuk memahami karakteristik awal *dataset*. Analisis mencakup distribusi kemunculan entitas, panjang teks, serta pola umum dalam judul berita. Tahap ini memastikan kualitas dan kesiapan data sebelum dilakukan *preprocessing*.

4. Tahap selanjutnya adalah *preprocessing* data, meliputi pembersihan teks (*cleaning*), normalisasi, dan tokenisasi. Proses *cleaning* menghapus karakter atau simbol yang tidak relevan, sedangkan normalisasi menyeragamkan format teks. Tokenisasi kemudian memecah teks menjadi unit *token* yang digunakan pada tahap pemrosesan linguistik.

5. *Dataset* yang telah dipreproses selanjutnya dibagi (data *splitting*) menjadi 80% data model dan 20% data *testing*. Data model tersebut kemudian dibagi kembali menjadi 80% data *training* dan 20% data *validation*. Pembagian ini dilakukan untuk memastikan proses pelatihan dan evaluasi model berlangsung objektif serta mencegah terjadinya data *leakage*.

6. Setelah pembagian data, dilakukan penandaan kelas kata (*Part-of-Speech*/POS *tagging*) menggunakan Stanza pada masing-masing subset data, yaitu *training*, *validation*, dan *testing*. POS *tagging* memberikan informasi kelas kata pada setiap *token* yang berfungsi sebagai fitur linguistik tambahan dalam proses pemodelan.

7. Tahap berikutnya adalah pembangunan *word embedding* menggunakan FastText untuk menghasilkan representasi vektor berbasis *subword*. *Embedding* ini menangkap karakteristik morfologis bahasa Indonesia dan diterapkan pada setiap subset data sehingga menghasilkan fitur numerik yang siap digunakan dalam model.

8. Setelah tahap *preprocessing*, POS *tagging*, dan pembentukan *word embedding* selesai pada masing-masing subset data (*training*, *validation*, dan *testing*), seluruh hasil pemrosesan disimpan dalam format `pickle` dan `JSON`. Penyimpanan ini dilakukan untuk memisahkan proses persiapan data dan proses pemodelan ke dalam dua *notebook* terpisah. *Notebook* pertama digunakan untuk menghasilkan serta menyimpan berkas `pickle` dan `JSON` yang memuat fitur serta label hasil pemrosesan, sedangkan *notebook* kedua

digunakan untuk memuat kembali berkas tersebut pada tahap pelatihan model. Pendekatan ini memastikan proses pemodelan dapat dilakukan tanpa perlu mengulangi tahap *preprocessing*, POS *tagging*, dan *embedding*, sehingga eksperimen berlangsung lebih efisien, terstruktur, dan mudah direproduksi.

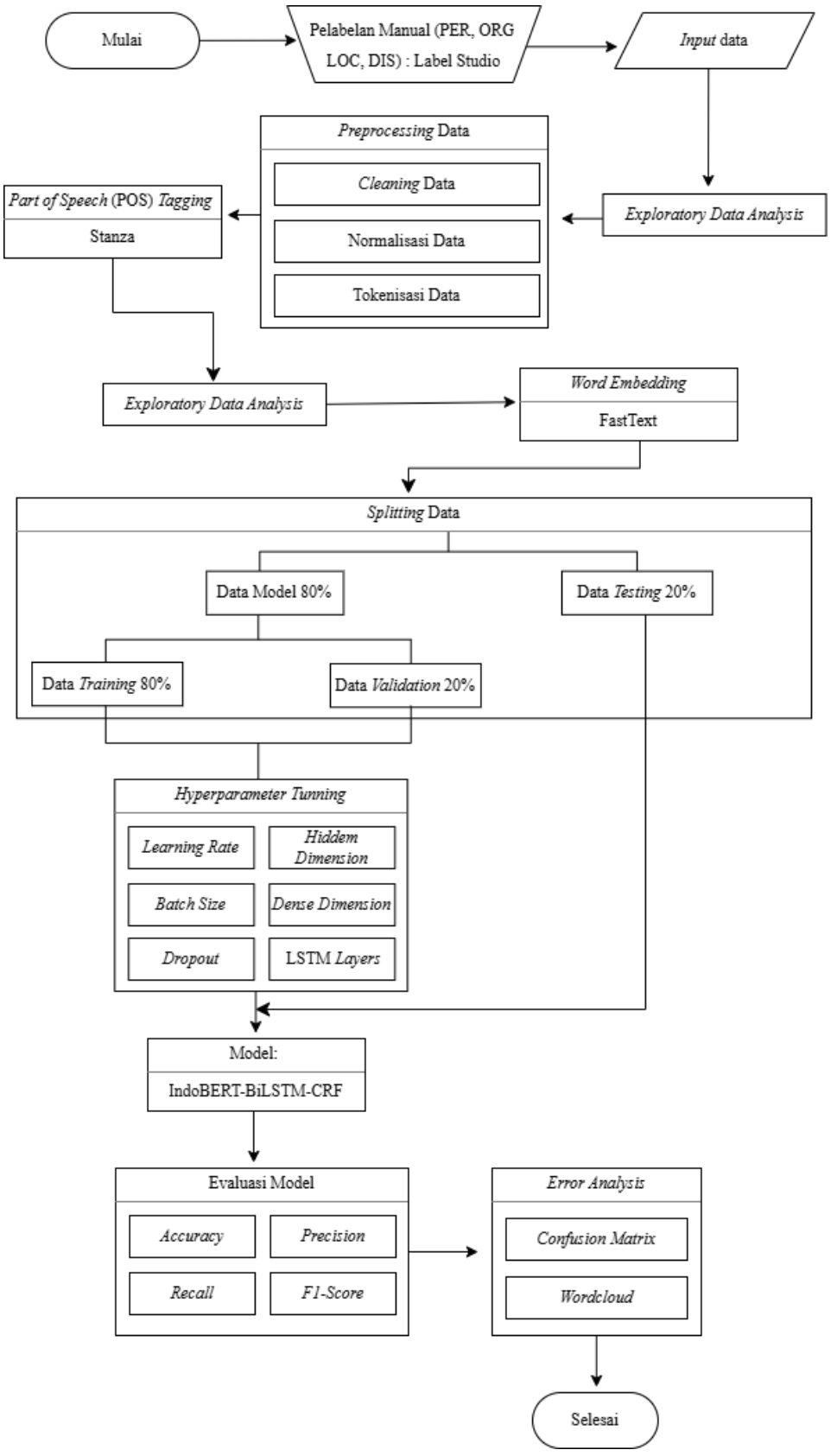
9. Selanjutnya dilakukan penyetelan hiperparameter (*hyperparameter tuning*) dengan memanfaatkan data *training* dan *validation*. Parameter yang dioptimasi meliputi *learning rate*, *batch size*, *dropout*, *hidden dimension*, *dense dimension*, serta jumlah lapisan LSTM. Tahap ini bertujuan memperoleh konfigurasi terbaik yang meningkatkan performa model sebelum evaluasi akhir.

10. Model *hybrid* IndoBERT–BiLSTM–CRF kemudian dibangun dengan mengintegrasikan kemampuan *contextual embedding* dari IndoBERT, kemampuan BiLSTM dalam menangkap dependensi sekuensial, serta CRF untuk meningkatkan konsistensi pelabelan sekuens. Model dilatih menggunakan data *training* dan dievaluasi secara berkala menggunakan data *validation* hingga diperoleh model terbaik.

11. Setelah diperoleh model terbaik, dilakukan evaluasi akhir menggunakan data *testing* yang tidak terlibat dalam proses pelatihan maupun tuning. Evaluasi dilakukan dengan metrik *accuracy*, *precision*, *recall*, dan F1-*score* untuk menilai kemampuan generalisasi model terhadap data baru.

12. Tahap selanjutnya adalah analisis kesalahan (*error analysis*) dengan menyusun *confusion matrix* dan *word cloud* guna mengidentifikasi pola kesalahan prediksi model, khususnya dalam mendeteksi entitas penyakit (DIS). Analisis ini memberikan gambaran mengenai sumber kesalahan serta kelemahan model dalam klasifikasi entitas.

13. Tahap akhir berupa penarikan kesimpulan yang mencakup konfigurasi *hyperparameter* optimal serta evaluasi efektivitas model *hybrid* IndoBERT–BiLSTM–CRF dalam mendeteksi entitas penyakit pada tugas NER.



Gambar 11. Langkah-langkah Penelitian.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan pada Bab IV mengenai pengembangan sistem *Named Entity Recognition* (NER) pada judul berita penyakit menular berbahasa Indonesia menggunakan model hibrida IndoBERT–BiLSTM–CRF dengan dukungan fitur POS Stanza dan FastText *embedding*, maka diperoleh beberapa kesimpulan sebagai berikut.

1. Model IndoBERT–BiLSTM–CRF yang dikombinasikan dengan POS Stanza dan FastText *embedding* berhasil dibangun dan menunjukkan kinerja yang sangat baik dalam mengenali entitas penyakit pada judul berita berbahasa Indonesia. Evaluasi pada data uji menghasilkan nilai *accuracy* sebesar 98,26%, F1-*micro* sebesar 0,9826, dan *weighted* F1-*score* sebesar 0,9828. Pada tingkat entitas (*entity-level*), model mencapai tingkat ketepatan sekitar 99% dalam mendeteksi entitas penyakit (DIS). Hasil ini menunjukkan bahwa integrasi representasi kontekstual (IndoBERT), sintaktik (POS *tagging*), dan leksikal (FastText) mampu memperkuat pemodelan sekuens dan meningkatkan konsistensi pelabelan BIO pada domain berita kesehatan.

2. Proses *hyperparameter tuning* menunjukkan bahwa kapasitas representasi sekuens pada lapisan BiLSTM berperan penting dalam peningkatan performa model. Konfigurasi terbaik diperoleh pada *hidden dimension* sebesar 256 dengan satu lapisan LSTM tanpa *dense layer* tambahan, *dropout* sebesar 0,22, *learning rate* sebesar $2{,}53 \times 10^{-5}$, dan *batch size* 16. Hasil ini menunjukkan bahwa peningkatan kapasitas representasi langsung pada lapisan BiLSTM lebih efektif dalam menangkap konteks sekuens dibandingkan penambahan lapisan proyeksi tambahan. Selain itu, kombinasi *learning rate* yang stabil

dan *dropout* moderat terbukti menjaga keseimbangan antara kemampuan generalisasi dan risiko *overfitting*.

3. Analisis kesalahan menunjukkan bahwa kelemahan utama model terletak pada prediksi label *inside* (I-*), khususnya I-DIS, yang memiliki distribusi data sangat terbatas. Model cenderung memprediksi *token* lanjutan dari entitas penyakit sebagai B-DIS atau O, sehingga konsistensi struktur BIO pada tingkat *token* belum sepenuhnya terjaga. Meskipun demikian, kesalahan ini tidak berdampak signifikan terhadap keberhasilan deteksi entitas penyakit secara utuh pada tingkat *entity-level*. Temuan ini menunjukkan bahwa permasalahan pada label I-DIS lebih disebabkan oleh ketidakseimbangan distribusi label dan keterbatasan variasi sekuens dalam data, bukan oleh kegagalan model dalam memahami konteks semantik entitas penyakit.

5.2 Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan, beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya adalah sebagai berikut.

1. Penelitian lanjutan disarankan untuk memperluas jumlah dan variasi data anotasi, khususnya pada entitas penyakit multi-*token* dengan label I-DIS, guna memperbaiki distribusi label dan meningkatkan stabilitas pembelajaran dalam struktur BIO. Selain itu, skema pelabelan alternatif seperti BIOES atau metode berbasis *span-level* dapat dipertimbangkan untuk mengurangi ketidakkonsistenan batas entitas. Pendekatan berbasis representasi topik atau struktur graf seperti *Latent Dirichlet Allocation* (LDA) atau *Graph Convolutional Network* (GCN) juga berpotensi digunakan untuk memodelkan hubungan kontekstual antar *token* sehingga dapat mengurangi kesalahan transisi dari label I-DIS ke O.

2. Strategi pelatihan yang lebih adaptif terhadap label minoritas dapat dipertimbangkan, misalnya melalui penyesuaian bobot kelas atau penerapan data *augmentation* berbasis linguistik. Pendekatan ini diharapkan dapat memperbaiki performa model pada label dengan distribusi terbatas tanpa menurunkan kinerja keseluruhan.

3. Penelitian berikutnya dapat memperluas cakupan evaluasi model pada skala data yang lebih besar dan beragam, termasuk teks berita lengkap, laporan epidemiologi, atau sumber media daring lainnya. Hal ini penting untuk menguji kemampuan generalisasi model dalam konteks yang lebih luas serta memastikan performa tinggi tidak hanya terbatas pada judul berita penyakit menular.

4. Pengembangan sistem ke arah implementasi praktis, seperti pembuatan prototipe aplikasi atau sistem pemantauan berbasis web, dapat menjadi langkah lanjutan yang relevan. Sistem tersebut dapat dimanfaatkan untuk mengekstraksi entitas penyakit secara otomatis dari berita daring dan mendukung pemantauan isu kesehatan masyarakat secara *real-time*.

5. Pengembangan skema anotasi yang lebih rinci juga dapat dipertimbangkan dengan menambahkan entitas lain yang relevan seperti lokasi (LOC), waktu kejadian (*DATE*), atau jumlah kasus (NUM). Ekspansi label ini akan memperkaya kemampuan model dalam melakukan ekstraksi informasi terstruktur serta meningkatkan potensi penerapan sistem dalam surveilans penyakit dan analisis epidemiologi berbasis teks.

DAFTAR PUSTAKA

Achonwa, E. C., & Adedeji, A. E. (2025). Speed vs. accuracy in digital journalism: A comparative analysis of Nigeria, the United States, and the United Kingdom. *International Journal of Advanced Multidisciplinary Research and Studies*, **5**(4), 1569–1574.

Akpinar, M. T. (2023). From organizational learning to machine learning with supervised, unsupervised, and reinforcement learning approach. Dalam M. Oduncuoglu & H. I. Kurt (Eds.), *Recent advances in natural and engineering sciences* (hlm. 155–168). Livre de Lyon.

Ali, Y. S., Searan, S. M., Qasim, R. H., & Aliesawi, S. A. (2025). Hyperparameter optimization for CNN, K-NN, and decision tree in handwritten digit classification. *Ingénierie des Systèmes d'Information*, **30**(5), 1201–1207. doi:10.18280/isi.300508

Anam, M., Defit, S., Haviluddin, H., Efrizoni, L., & Firdaus, M. (2024). Early stopping on CNN-LSTM development to improve classification performance. *Journal of Applied Data Sciences*, **5**(3), 1175–1188. doi:10.47738/jads.v5i3.312

Asosiasi Penyelenggara Jasa Internet Indonesia. (2025). *Survei Internet APJII 2025*.

Asri, Y., & Kuswardani, D. (2024). *Machine Learning & Deep Learning: Analisis Sentimen Menggunakan Ulasan Pengguna Aplikasi*. Uwais Inspirasi Indonesia.

Aung, N. N., Pang, J., Chua, M. C. H., & Tan, H. X. (2023). A novel bidirectional LSTM deep learning approach for COVID-19 forecasting. *Scientific Reports*, **13**(1), 17953. doi:10.1038/s41598-023-44924-8

Bhadauria, D., Sierra-Múnera, A., & Krestel, R. (2024). The effects of data quality on named entity recognition. Dalam *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)* (hlm. 79–88). Association for Computational Linguistics.

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, **9**(1), 10. doi:10.1186/s40537-022-00561-y

Dave, E., & Chowanda, A. (2024). IPerFEX-2023: Indonesian personal financial entity extraction using indoBERT-BiGRU-CRF model. *Journal of Big Data*, **11**(1), Article 139. doi:10.1186/s40537-024-00987-6

Fakih, A., Hamzami, M. A., Hadianto, M. R., & Alifah, N. I. S. (2025). Perbandingan akurasi algoritma C4.5 dan K-NN untuk prediksi kelulusan mahasiswa penerima beasiswa. *Jurnal Komputer Antartika*, **3**(1), 18–25. https://doi.org/10.70052/jka.v3i1.623

Finansyah, A. Y. W., Afiahayati, F., & Sutanto, V. M. (2022). Performance comparison of similarity measure algorithm as data preprocessing stage: Text normalization in Bahasa. *Scientific Journal of Informatics*, **9**(1), 1–7. doi:10.15294/sji.v9i1.30052

Gajiwala, C. (2025). The rise of deep learning and neural networks: Revolutionizing artificial intelligence. *European Journal of Computer Science and Information Technology*, **13**(17), 88–98. doi:10.37745/ejcsit.2013/vol13n178898

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*. doi:10.48550/arXiv.2008.05756

Gustineli, M. (2022). A survey on recently proposed activation functions for deep learning. *arXiv preprint arXiv:2204.02921*. doi:10.48550/arXiv.2204.02921

Hou, Y., & Huang, J. (2025). Natural language processing for social science research: A comprehensive review. *Chinese Journal of Sociology*, **11**(1), 121–157. doi:10.1177/2057150X241306780

Ibrahim, V., Abu Bakar, J., Harun, N. H., & Abdulateef, A. F. (2021). A word cloud model based on hate speech in an online social media environment. *Baghdad Science Journal*, **18**(2 Suppl.), 45. doi:10.21123/bsj.2021.18.2(Suppl.).0937

Jarrar, M., Khalilia, M., & Ghanem, S. (2022). Wojood: Nested Arabic named entity corpus and recognition using BERT. *arXiv preprint arXiv:2205.09651*. doi:10.48550/arXiv.2205.09651

Jiang, H., Hua, Y., Beeferman, D., & Roy, D. (2022). Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. *arXiv preprint arXiv:2201.07281*. doi:10.48550/arXiv.2201.07281

Julianto, A., Sunyoto, A., & Wibowo, F. W. (2022). Optimasi hyperparameter convolutional neural network untuk klasifikasi penyakit tanaman padi. *TEKNIMEDIA: Teknologi Informasi dan Multimedia*, **3**(2), 98–105. doi:10.46764/teknimedia.v3i2.77

Kahloot, K. M., & Ekler, P. (2021). Algorithmic splitting: A method for dataset preparation. *IEEE Access*, **9**, 125229–125237. doi:10.1109/ACCESS.2021.3110745

Karo, I. M. K., Dewi, S., & Syahrin, A. (2025). Ekstraksi informasi bencana banjir dari berita online berbasis named entity recognition. *MULTINETICS*, **11**(1), 35–42. doi:10.32722/multinetics.v11i1.7499

Keerthana, R., & Uddin, W. (2024). Named entity recognition using NLP. *International Journal of Science, Engineering and Technology*, **12**(2). doi:10.61463/ijset.vol.12.issue2.142

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*. doi:10.48550/arXiv.2401.10825

Ketmaneechairat, H., & Maliyaem, M. (2020). Natural language processing for disaster management using conditional random fields. *Journal of Advances in Information Technology*, **11**(2), 97–102. doi:10.12720/jait.11.2.97-102

Khairunnisa, S. O., Chen, Z., & Komachi, M. (2023). Dataset enhancement and multilingual transfer for named entity recognition in the Indonesian language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **22**(6), 1–21. doi:10.1145/3592854

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *arXiv preprint arXiv:2011.00677*. doi:10.48550/arXiv.2011.00677

Kurniasari, D., Pradana, F. A., & Wamiliana, W. (2025). Deep learning-driven comparative study of word embedding techniques: Word2Vec, GloVe, and FastText

in health condition reviews. *ISRG Journal of Multidisciplinary Studies*, **3**(2), 51–62. doi:10.5281/zenodo.14890438

Kurniawan, K., Ceasaro, B., & Sucipto, S. (2024). Perbandingan fungsi aktivasi untuk meningkatkan kinerja model LSTM dalam prediksi ketinggian air sungai. *JEPIN*, **10**(1), 134–143. doi:10.26418/jp.v10i1.72866

Lamprou, Z., Polick, F., & Moshfeghi, Y. (2025). Aligning brain activity with advanced transformer models: Exploring the role of punctuation in semantic processing. *arXiv preprint arXiv:2501.06278*. doi:10.48550/arXiv.2501.06278

Li, H., Rajbahadur, G. K., Lin, D., Bezemer, C.-P., & Jiang, Z. M. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. *IEEE Access*, **12**, 70676–70689. doi:10.1109/ACCESS.2024.3402543

Liu, Z., Xu, Z., Jin, J., Shen, Z., & Darrell, T. (2023). Dropout reduces underfitting. In *International Conference on Machine Learning (ICML)* (hlm. 22233–22248). PMLR. doi:10.48550/arXiv.2303.01500

Manurung, M. T., Wijayakusuma, I. G. N. L., & Gautama, I. P. W. (2025). Named entity recognition for medical records of heart failure using a pre-trained BERT model. *Journal of Applied Informatics and Computing*, **9**(2), 341–348. doi:10.30871/jaic.v9i2.9170

Maurya, M. (2023). Name entity recognition and various tagging schemes. *Medium*.

Montesinos-López, O. A., Montesinos, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction*. Springer Nature. doi:10.1007/978-3-030-89010-0

Muraina, I. O. (2022). Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts. Dalam *Proceedings of the 7th International Mardin Artuklu Scientific Research Conference* (hlm. 496–504).

Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical & Computer Engineering*, **14**(1), 1071–1078. doi:10.11591/ijece.v14i1.pp1071-1078

Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2023). BERT base model for toxic comment analysis on Indonesian social media. *Procedia Computer Science*, **216**, 714–721. doi:10.1016/j.procs.2022.12.188

Natzir, S. M. (2023). Perbandingan kinerja model pembelajaran mesin dalam prediksi banjir menggunakan KNN, Naive Bayes, dan Random Forest. *HOAQ (High Education of Organization Archive Quality): Jurnal Teknologi Informasi*, **14**(2), 59–64. https://doi.org/10.52972/hoaq.vol14no2.p59-64

Nurhakiki, J., & Yahfizham, Y. (2024). Studi kepustakaan: Pengenalan 4 algoritma pada pembelajaran deep learning beserta implikasinya. *Pendekar: Jurnal Pendidikan Berkarakter*, **2**(1), 270–281. doi:10.51903/pendekar.v2i1.598

Pakhale, K. (2023). Comprehensive overview of named entity recognition: Models, domain specific applications and challenges. *arXiv preprint arXiv:2309.14084*. doi:10.48550/arXiv.2309.14084

Patil, N., Patil, A., & Pawar, B. V. (2020). Named entity recognition using conditional random fields. *Procedia Computer Science*, **167**, 1181–1188. doi:10.1016/j.procs.2020.03.431

Perdana, R. S., & Adikara, P. P. (2025). Multi-task learning for named entity recognition and intent classification in natural language understanding applications. *Journal of Information Systems Engineering and Business Intelligence*, **11**(1), 1–16. doi:10.20473/jisebi.11.1.1-16

Purwitasari, N. A., & Soleh, M. (2022). Implementasi algoritma artificial neural network dalam pembuatan chatbot menggunakan pendekatan natural language processing. *Jurnal Ilmu Pengetahuan dan Teknologi*, **6**(1), 14–21. doi:10.31543/jii.v6i1.192

Puspitasari, A., Paradhita, A. N., Tineka, Y. W., Sulistyowati, V., & Noriska, N. K. S. (2024). Natural language processing (NLP) technology for chatbot website. *Jurnal Penelitian Pendidikan IPA*, **10**(Special Issue), 319–324. doi:10.29303/jppipa.v10iSpecialIssue.8241

Putra, A., & Kurniawan, R. (2021). Bidirectional LSTM-CNNs untuk ekstraksi entitas lokasi kebakaran pada berita online berbahasa Indonesia. *Seminar Nasional Official Statistics*, **2020**(1), 319–327. doi:10.34123/semnasoffstat.v2020i1.601

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*. doi:10.48550/arXiv.2003.07082

Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, **35**(23), 17095–17112. doi:10.1007/s00521-023-08568-z

Ryu, S., Chun, J. Y., Lee, S., Yoo, D., Kim, Y., Ali, S. T., & Chun, B. C. (2022). Epidemiology and transmission dynamics of infectious diseases and control measures. *Viruses*, **14**(11), 2510. doi:10.3390/v14112510

Sathyanarayanan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, **27**(4S), 4023–4031. doi:10.53555/AJBR.v27i4S.4345

Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. *arXiv preprint arXiv:2305.17473*. doi:10.48550/arXiv.2305.17473

Siregar, P. S., Hatika, R. G., & Hayadi, B. H. (2023). Multiple choice question difficulty level classification with multi class confusion matrix in the online question bank of Education Gallery. *Journal of Applied Data Sciences*, **4**(4), 392–406. doi:10.47738/jads.v4i4.132

Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics: A review. *Cognitive Robotics*, **3**, 54–70. doi:10.1016/j.cogr.2023.04.001

Sujadi, H. (2022). Analisis sentimen pengguna media sosial Twitter terhadap wabah COVID-19 dengan metode Naive Bayes Classifier dan Support Vector Machine. *INFOTECH Journal*, **8**(1), 22–27. doi:10.31949/infotech.v8i1.1883

Turuta, O., & Maksymenko, D. (2025). Tokenization efficiency of current foundational large language models for the Ukrainian language. *Frontiers in Artificial Intelligence*, **8**, 1538165. doi:10.3389/frai.2025.1538165

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. *arXiv preprint arXiv:2001.09636*. doi:10.48550/arXiv.2001.09636

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. doi:10.48550/arXiv.1706.03762

Warto, Rustad, S., Shidik, G. F., Noersasongko, E., Purwanto, Muljono, & Setiadi, D. R. I. M. (2024). Systematic literature review on named entity recognition: Approach, method, and application. *Statistics, Optimization & Information Computing*, **12**(4), 907–942. doi:10.19139/soic-2310-5070-1631

Wei, X., Salsabil, L., & Wu, J. (2022). Theory entity extraction for social and behavioral sciences papers using distant supervision. Dalam *Proceedings of the 22nd ACM Symposium on Document Engineering* (hlm. 1–4). Association for Computing Machinery. doi:10.1145/3558100.3563855

Westlund, O., Boyles, J. L., Guo, L., Saldaña, M., Salaverría, R., Thomson, T. J., & Wu, S. (2025). Digital journalism (studies): An agenda for the future. *Digital Journalism*, **13**(2), 179–194. doi:10.1080/21670811.2025.2474530

Williams, C., & Archibong, B. (2024). The roles of mass media in information dissemination: The process and challenges. Global Academic Star Publications.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*. doi:10.48550/arXiv.2009.05387

Wiujianna, A., Sunarno, S., & Iqbal, I. (2025). Perbandingan performa model Long Short-Term Memory dan Bidirectional untuk prediksi kabut. *Jurnal Teknik Informatika dan Sistem Informasi*, **11**(2), 251–260. doi:10.28932/jutisi.v11i2.10588

Wszola, E., Jaggi, M., & Püschel, M. (2021). Faster parallel training of word embeddings. Dalam *2021 IEEE 28th International Conference on High Performance Computing, Data, and Analytics (HiPC)* (hlm. 31–41). IEEE. doi:10.1109/HiPC53243.2021.00017

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, **415**, 295–316. doi:10.1016/j.neucom.2020.07.061

Yulianti, E., Bhary, N., Abdurrohman, J., Dwitilas, F. W., Nuranti, E. Q., & Husin, H. S. (2024). Named entity recognition on Indonesian legal documents: A dataset and study using transformer-based models. *International Journal of Electrical and Computer Engineering*, **14**(5), 5489–5501. doi:10.11591/ijece.v14i5.pp5489-5501

Yulianti, E., & Nissa, N. K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches. *Bulletin of Electrical Engineering and Informatics*, **13**(5), 3579–3589. doi:10.11591/eei.v13i5.8032

Zhou, F., Hou, F., Wang, J., Ma, Q., & Luo, L. (2024). Prevention and control of infectious disease transmission in subways: An improved susceptible–exposed–infected–recovered model. *Frontiers in Public Health*, **12**, 1454450. doi:10.3389/fpubh.2024.1454450