IMPLEMENTASI *NAMED ENTITY RECOGNITION* (NER) BERBASIS *DEEP LEARNING* UNTUK EKSTRAKSI INFORMASI PADA BERITA ONLINE MENGENAI PENYAKIT MENULAR

(Skripsi)

Oleh

ANASTASIA HARUM MAWADAH
2117031107

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRACT

IMPLEMENTATION OF NAMED ENTITY RECOGNITION (NER) BASED ON DEEP LEARNING FOR INFORMATION EXTRACTION IN ONLINE NEWS ABOUT INFECTIOUS DISEASES

By

Anastasia Harum Mawadah

Infectious diseases remain a major public health challenge in Indonesia, with case trends continuing to rise. In the digital era, although information spreads rapidly through online media, the main challenge lies in extracting relevant data from the vast volume of available information. This study aims to apply the Named Entity Recognition (NER) method with the addition of Part-of-Speech (POS) Tagging and to develop BiLSTM, BiLSTM-CRF, and hybrid IndoBERT-BiLSTM-CRF models to extract information from online news related to infectious diseases. The results show that the NER method successfully identified and labeled entities such as locations, organizations, and individuals, making the information extraction process more efficient and systematic. The best-performing model was the hybrid combination of IndoBERT and BiLSTM-CRF, with an 80% training and 20% testing data split, yielding optimal accuracy, precision, recall, and F1-score values.

Keywords: Named Entity Recognition (NER), Part-of-Speech (POS) Tagging, BiLSTM, CRF, IndoBERT, Information Extraction, Online News, Infectious Diseases.

ABSTRAK

IMPLEMENTASI *NAMED ENTITY RECOGNITION* (NER) BERBASIS *DEEP LEARNING* UNTUK EKSTRAKSI INFORMASI PADA BERITA ONLINE MENGENAI PENYAKIT MENULAR

Oleh

Anastasia Harum Mawadah

Penyakit menular menjadi tantangan besar bagi kesehatan masyarakat di Indonesia dengan tren kasus yang terus meningkat. Di era digital, meskipun informasi tersebar cepat melalui media online, tantangan utama adalah mengekstraksi data yang relevan dari volume informasi yang sangat besar. Penelitian ini bertujuan menerapkan metode NER dengan penambahan POS Tagging dan membangun model BiLSTM, BiLSTM-CRF, serta IndoBERT-BiLSTM-CRF untuk mengekstraksi informasi dari berita online terkait penyakit menular. Hasil penelitian menunjukkan bahwa metode NER berhasil mengenali dan memberi label pada entitas seperti lokasi, organisasi, dan tokoh, sehingga proses ekstraksi informasi menjadi lebih efisien dan sistematis. Model terbaik diperoleh dari kombinasi IndoBERT dan BiLSTM-CRF dengan akurasi sebesar 95%. Pembagian data 80% untuk pelatihan dan 20% untuk pengujian, menghasilkan nilai akurasi, presisi, recall, dan f1-score yang optimal.

Kata-kata kunci: Penyakit Menular, Named Entity Recognition (NER), Part Of Speech (POS) Tagging, BiLSTM, CRF, IndoBERT, Berita Online, Ekstraksi Informasi.

IMPLEMENTASI *NAMED ENTITY RECOGNITION* (NER) BERBASIS *DEEP LEARNING* UNTUK EKSTRAKSI INFORMASI PADA BERITA ONLINE MENGENAI PENYAKIT MENULAR

Oleh

ANASTASIA HARUM MAWADAH

(Skripsi)

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

Judul Skripsi	:	**IMPLEMENTASI *NAMED ENTITY RECOGNATION* (NER) BERBASIS *DEEP LEARNING* UNTUK EKTRAKSI INFORMASI PADA BERITA ONLINE MENGENAI PENYAKIT MENULAR**
Nama Mahasiswa	:	*Anastasia Harum Mawadah*
Nomor Pokok Mahasiswa	:	**2117031107**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Dian Kurniasari, S.Si., M.Sc.
NIP. 196903051996032001

Dr. Purnomo Husnul Khotimah, M.T.
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 : **Ir. Warsono, M.S., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **19 Juni 2025**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Anastasia Harum Mawadah**
Nomor Pokok Mahasiswa	:	**2117031107**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Implementasi *Named Entity Recognation* (NER) Berbasis *Deep Learning* Untuk Ekstraksi Informasi Pada Berita Online Mengenai Penyakit Menular**

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, 19 Juni 2025
Penulis,

Anastasia Harum Mawadah
NPM. 2117031107

RIWAYAT HIDUP

Penulis bernama lengkap Anastasia Harum Mawadah, lahir di Kota Bekasi, Jawa Barat, pada tanggal 13 Juli 2002. Penulis merupakan anak kedua dari dua bersaudara, putri dari pasangan Bapak Muklis Ahmadi dan Ibu Nasroh.

Penulis mengawali pendidikan di Taman Kanak-Kanak (TK) Nurul Hikmah pada tahun 2007–2008, kemudian melanjutkan pendidikan dasar di SDN Mustika Jaya VI Kota Bekasi pada tahun 2008–2014. Selanjutnya, penulis menempuh pendidikan menengah pertama di SMP Negeri 5 Pesawaran pada tahun 2014–2017, dan melanjutkan ke jenjang menengah atas di SMA Negeri 1 Gedongtataan pada tahun 2017–2020. Penulis diterima sebagai mahasiswi Program Studi S1 Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung pada tahun 2021.

Selama menjalani masa studi, penulis aktif dalam berbagai kegiatan organisasi. Penulis pernah menjabat sebagai Sekretaris Biro Kemuslimahan dalam kepengurusan ROIS FMIPA Universitas Lampung periode 2023–2024, serta menjadi Staf Ahli di Badan Eksekutif Mahasiswa (BEM) FMIPA Universitas Lampung pada Dinas Sosial dan Pengabdian Masyarakat (SPM). Selain itu, penulis juga aktif sebagai anggota bidang *Public Relations and Marketing* (PRM) dalam organisasi Volunteer *World Cleanup Day* (WCD) Lampung untuk periode 2023–2025.

Selanjutnya, pada bulan Desember 2023 hingga Juni 2024, penulis melaksanakan Praktik Kerja Lapangan (PKL) yang dilanjutkan dengan mengikuti program Merdeka Belajar Kampus Merdeka (MBKM) di Badan Riset dan Inovasi Nasional (BRIN) yang berlokasi di Bandung. Kemudian, pada bulan September hingga Desember 2024, penulis mengikuti program Studi Independen Bersertifikat melalui program BANGKIT yang diselenggarakan oleh PT Dicoding Akademi Indonesia sebagai *cohort Machine Learning*.

KATA INSPIRASI

”*Lakukan yang bisa dikendalikan, terima yang tidak bisa diubah*.”
(Penulis)

”*Angan-angan yang dulu mimpi belaka. Kita gapai segala yang tak disangka*.”
(Hindia)

”*Maka, sesungguhnya bersama kesulitan ada kemudahan.*
Sesungguhnya beserta kesulitan ada kemudahan.”
(QS. Al-Insyirah: 5-6)

”*Boleh jadi kamu membeci sesuatu, padahal ia amat baik bagi kamu, dan boleh jadi kamu mencintai sesuatu, padahal ia amat buruk bagi kamu. Allah maha mengetahui sedangkan kamu tidak mengetahui*.”
(QS. Al-Baqarah: 216)

PERSEMBAHAN

Dengan mengucap Alhamdulillah dan syukur kepada Allah SWT atas nikmat serta hidayah-Nya sehingga skripsi ini dapat terselesaikan dengan baik dan tepat pada waktunya. Dengan rasa syukur dan Bahagia, saya persembahkan rasa terimakasih saya kepada:

Ayah dan Ibuku Tercinta

Terimakasih kepada orang tuaku atas segala pengorbanan, motivasi, doa dan ridho serta dukungannya selama ini. Terimakasih telah memberikan pelajaran berharga kepada anakmu ini tentang makna perjalanan hidup yang sebenarnya sehingga kelak bisa menjadi orang yang bermanfaat bagi banyak orang.

Dosen Pembimbing dan Pembahas

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

Sahabat-sahabatku

Terimakasih kepada semua orang-orang baik yang telah memberikan pengalaman, semangat, motivasinya, serta doa-doanya dan senantiasa memberikan dukungan dalam hal apapun.

Almamater Tercinta
Universitas Lampung

SANWACANA

Alhamdulillah, puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini yang berjudul ”Implementasi *Named Entity Recognition* (NER) Berbasis *Deep Learning* untuk Ekstraksi Informasi pada Berita Online Mengenai Penyakit Menular” dengan baik dan lancar serta tepat pada waktu yang telah ditentukan. Shalawat serta salam semoga senantiasa tercurahkan kepada Nabi Muhammad SAW.

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 1 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. Bapak Ir. Warsono, M.S., Ph.D. selaku Pembahas yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis sehingga dapat menjadi lebih baik lagi.
4. Bapak Andri Fachrur Rozie, S.Kom., M.Eng. selaku salah satu pembimbing MBKM BRIN KST Samaun Samadikun, Bandung.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Ibu Dra. Dorrah Aziz, M.Si., Ibu Dr. Khoirin Nisa, S.Si., M.Si., serta Bapak Agus Sutrisno, S.Si., M.Si. selaku dosen yang senantiasa memberikan semangat, motivasi, serta dukungan selama proses penyusunan skripsi.

8. Seluruh dosen, staff dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
9. Teristimewa untuk Kedua Orangtua tercinta, Bapak Muklis Ahmadi dan Mama Nasroh serta abang Ahmad Fajar Rizky Firdaus, S.Kom., yang telah memberikan kasih sayang, nasihat, motivasi, memenuhi kebutuhan, serta doa tiada henti kepada penulis selama proses perkuliahan.
10. Anak-anak Ibu Dian dan Pak War selaku teman seperbimbingan yang telah sama-sama berjuang dan saling menyemangati satu sama lain.
11. Sahabat seperjuangan Nofa dan Yulina yang telah membantu penulis dalam menjalani perkuliahan.
12. Teman-teman Jurusan Matematika Angkatan 2021.

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, 19 Juni 2025

Anastasia Harum Mawadah

DAFTAR ISI

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Penyakit menular menjadi tantangan kesehatan masyarakat di seluruh dunia, termasuk di Indonesia. Selama beberapa tahun terakhir, jumlah kasus penyakit menular menunjukkan tren peningkatan (Lestari dkk., 2021). Peningkatan ini mencerminkan masalah yang lebih besar dalam sistem kesehatan. Secara keseluruhan, total kasus penyakit menular di Indonesia melebihi 2,4 juta. (Kementerian Kesehatan Republik Indonesia, 2023). Kondisi ini terdeteksi di daerah dengan kepadatan penduduk tinggi dan infrastruktur kesehatan yang terbatas (Eckert dkk., 2014).

Penyebaran informasi yang akurat sangat diperlukan. Seiring dengan pesatnya perkembangan era digital, informasi mengenai kesehatan dan penyakit menular tersebut tersebar dengan cepat melalui berbagai platform media online. Artikel berita memainkan peran penting dalam menyebarkan informasi. Aksesibilitas dan kecepatan dalam penyebarannya memudahkan masyarakat untuk terus mengikuti perkembangan situasi krisis secara *real-time* (Kormelink dan Meijer, 2018). Namun, volume informasi yang besar ini juga membawa tantangan tersendiri. Tantangan utama yang dihadapi adalah ekstraksi data yang relevan.

Named Entity Recognition (NER) sebagai salah satu solusi dalam menghadapi tantangan ini. Metode NER untuk pertama kalinya diperkenalkan di *Message Understanding Conference-6* (Grishman dan Sundheim, 1996). Komponen utama dari NER bertujuan untuk mendeteksi dan mengklasifikasikan entitas bernama dalam suatu teks. Metode NER umumnya digunakan untuk mengenali nama orang, lokasi, dan organisasi dalam sebuah dokumen, namun dapt disesuikan. Selain itu, NER juga merupakan komponen penting yang mendasari banyak aplikasi dalam *Natural Language Processing* (NLP) (Yadav dan Bethard, 2019).

Penggunaan NER banyak digunakan dengan berbagai entitas. Sebagian besar penelitian NER hingga saat ini telah memberikan perhatian utama pada kategori penting yang dikenal sebagai *Big Three*, yaitu *person* (tokoh), *organization* (organisasi), dan *location* (lokasi) (Nadeau dan Sekine, 2007). Penggabungan ketiga kategori justru membuka peluang untuk eksplorasi yang mendalam terhadap masing-masing jenis entitas secara spesifik. Fokus pada pelabelan entitas lokasi sangat penting dalam konteks pemantauan dan penanggulangan wabah penyakit karena dapat membantu mengidentifikasi pola penyebaran secara geografis, mempercepat respons terhadap situasi darurat, serta mendukung pengambilan keputusan berbasis wilayah. Di sisi lain, pelabelan entitas orang memungkinkan pengenalan individu yang terlibat atau terdampak, seperti pasien dan tenaga medis, sementara pelabelan organisasi dapat mengungkap peran instansi atau otoritas dalam penanganan kasus dan kebijakan kesehatan. Identifikasi entitas yang tepat dan terstruktur, analisis data menjadi lebih komprehensif, akurat, dan dapat memberikan landasan bagi pengambilan kebijakan yang efektif.

Selain itu, pengembangan model NER yang fokus pada artikel berita online, terutama pada judul berita dimana biasanya berisi ringkasan dan mewakili isi konten pada artikel (Khotimah dkk., 2023) menjadi penting karena adanya tantangan unik terkait dengan konteks wabah penyakit (Zhang dkk., 2024). Meskipun teks formal seperti judul pada artikel berita online umumnya memiliki format yang terstruktur, serta memudahkan proses pengenalan entitas namun masih ditemukan tantangan dalam bentuk penamaan yang ambigu, tidak lengkap, atau kurang spesifik terhadap konteks. Hal ini dapat memengaruhi akurasi dalam proses ekstraksi informasi, terutama ketika nama tempat, tokoh, atau institusi disampaikan secara tidak eksplisit. Seiring dengan kemudahan akses data. Pembelajaran mesin memungkinkan pengembangan model yang lebih canggih, di mana model dibangun berdasarkan data yang telah dilabeli sebelumnya.

Bahasa Indonesia merupakan bahasa yang dituturkan oleh hampir 200 juta orang dan merupakan bahasa ke-10 yang paling banyak dituturkan di dunia, namun bahasa ini kurang terwakili dalam penelitian NLP (Koto dkk., 2020). Penggunaan bahasa Indonesia pada penelitian NER masih menggunakan dataset yang terbatas dan belum mempertimbangkan variasi dialek bahasa Indonesia (Aji dkk., 2022). Tantangan utama dalam pengembangan NER untuk bahasa Indonesia adalah kurangnya dataset terstandarisasi serta kompleksitas struktur bahasa yang berbeda dari bahasa Inggris (Mahendra dkk., 2019).

Penggabungan penandaan *Part Of Speech* (POS) *Tagging* sebagai fitur model dapat diselidiki untuk meningkatkan kinerja tugas ekstraksi pada dataset yang sedikit yaitu proses menandai kata dalam teks dengan kategori gramatikal tertentu (seperti kata benda, kata kerja, kata sifat, dll) berdasarkan definisi dan konteksnya (Purnamasari dan Suwardi, 2018). Beberapa penelitian terdahulu telah memberikan kontribusi signifikan, yaitu berhasil mengembangkan metode-metode NER yang semakin canggih dan akurat. Memanfaatkan teknik pembelajaran mendalam dan model berbasis *transformer*.

Penelitian mengenai ekstraksi lokasi dari teks terkait lalu lintas telah dilakukan oleh Andri Fachrur Rozie dkk (2023). Penelitian tersebut mengusulkan pendekatan *hybrid* yang menggabungkan model *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Conditional Random Field* (CRF). Model yang diusulkan menggunakan POS *tagger* sebagai fitur tambahan untuk meningkatkan kinerja dalam pengenalan entitas. Tujuan utama dari penelitian ini adalah untuk mengidentifikasi lokasi dalam teks berbahasa Indonesia yang berkaitan dengan kejadian lalu lintas. Dataset yang digunakan berasal dari portal Twitter ”LewatMana.com” yang membahas tentang kejadian lalu lintas. Model *hybrid* yang diusulkan mengkombinasikan kekuatan BiLSTM dalam memahami urutan data serta kemampuan CRF dalam *sequence labeling*. Proses evaluasi dilakukan dengan menguji model menggunakan dataset yang telah dianotasi secara manual. Model berhasil mencapai akurasi 91.21% dalam mengekstrak entitas lokasi. Hasil ini menunjukkan bahwa kombinasi BiLSTM-CRF dengan POS *Tagging* serta optimasi parameter memberikan kinerja yang sangat baik dalam tugas ekstraksi lokasi dari teks sosial media berbahasa Indonesia.

Implementasi model *deep learning* berbasis *transformer*, seperti *Bidirectional Encoder Representations from Transformers* (BERT), telah dilakukan juga pada bahasa Indonesia. Wilie dkk (2020) membandingkan model BERT multilingual dengan *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT) menggunakan data berbahasa Indonesia. Pada penelitiannya model IndoBERT mencapai akurasi 88% dalam pengenalan entitas dibandingkan model BERT multilingual hanya mencapai akurasi 75%. Peningkatan akurasi ini menunjukkan kemampuan IndoBERT dalam memahami konteks bahasa Indonesia dengan lebih baik karena dilatih khusus menggunakan data berbahasa Indonesia, memungkinkan pemahaman yang lebih mendalam. Selain akurasi, IndoBERT juga menunjukkan kinerja yang lebih baik dalam metrik lainnya.

Penelitian Nazar dkk. (2021) menunjukkan bahwa model *Named Entity Recognition* (NER) tradisional dapat memberikan hasil yang lebih stabil dan konsisten, terutama ketika berhadapan dengan jumlah data yang terbatas. Namun, penelitian tersebut juga mengungkap bahwa penggunaan banyak label dalam NER menyebabkan penurunan akurasi, dengan nilai F1-*score* menurun dari 89,5% (dengan 4 kategori) menjadi 77,2% (dengan 15 kategori), di mana 30% kesalahan klasifikasi disebabkan oleh *over-labeling*, dan kompleksitas yang meningkat membuat model kesulitan membedakan entitas yang mirip. Hal ini menunjukkan bahwa peningkatan jumlah label tidak selalu berdampak positif terhadap performa model, terutama jika tidak disertai dengan strategi pelabelan yang tepat dan data yang mencukupi. Ketidakseimbangan distribusi label juga menjadi faktor utama yang menyebabkan model lebih sering keliru dalam mengklasifikasikan entitas minor.

Berdasarkan beberapa penelitian sebelumnya, penelitian ini bertujuan untuk mengumpulkan informasi entitas penting dari teks berbahasa Indonesia dengan fokus pada ekstraksi entitas bernama, yang mencakup lokasi (spasial), person (sosial), dan organization (institusional). Teks yang dianalisis berasal dari artikel berita online terkait penyakit menular yang diperoleh dari dua portal berita di Indonesia. Teknik pengenalan entitas bernama digunakan untuk mengidentifikasi dan mengelompokkan entitas-entitas ini dalam teks. Mengeksplorasi efektivitas arsitektur *deep learning*, meliputi BiLSTM, *Hybrid* BiLSTM-CRF, dan *Hybrid* IndoBERT-BiLSTM-CRF, serta penggabungan penandaan POS *tagging*. Oleh karena itu penelitian ini membahas ”Implementasi *Named Entity Recognition* (NER) Berbasis *Deep Learning* untuk Ekstraksi Informasi pada Berita Online Mengenai Penyakit Menular”.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, adapun rumusan masalah dalam penelitian ini diantaranya:

1. Penerapan pelabelan entitas menggunakan metode NER dalam mengekstraksi informasi dari berita online tentang penyakit menular.
2. Menghitung kinerja model BiLSTM, *Hybrid* BiLSTM-CRF, dan *Hybrid* IndoBERT-BiLSTM-CRF dalam mengekstraksi informasi dari berita online tentang penyakit menular.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menerapkan metode NER dalam mengekstraksi informasi dari berita online tentang penyakit menular.
2. Membangun model BiLSTM, *Hybrid* BiLSTM-CRF, dan *Hybrid* IndoBERT-BiLSTM-CRF serta mengevaluasi kinerja masing-masing model dalam meningkatkan akurasi ektraksi entitas lokasi berita online tentang penyakit menular.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah:

1. Memberikan pemahaman yang lebih baik mengenai penerapan metode NER dalam mengekstraksi informasi berita online tentang penyakit menular, serta untuk mengidentifikasi model yang paling efektif dalam meningkatkan akurasi ekstraksi entitas berita online.
2. Menjadi referensi bagi peneliti selanjutnya dalam melakukan pengklasifikasian entitas bernama menggunakan model yang sama dengan jenis data yang berbeda.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Pada subbab ini diberikan beberapa penelitian yang telah dilakukan oleh peneliti sebelummnya dan dijadikan refrensi pada penelitian ini. Penelitian diringkas dalam Tabel 1 sebagai berikut:

Tabel 1. Penelitian Terkait

No	Judul Penelitian	Data	Metode	Hasil
1	*Location Extraction from Traffic Event-related Text* (Andi dkk., 2023)	Channel Twitter "Lewat Mana.com"	BiLSTM-CRF dengan Indonesian CRF-Tagger	Akurasi dengan epoch 500 sebesar 91,01% sedangkan dengan epoch 200 sebesar 91,21%
2	*Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language* (Siti Oryza Khairunnisa, 2021)	*Monolingual Indonesian NER, Experiment Settings*	BERT, IndoBERT, mBERT, XLM-RoBERTa, BiLSTM-CRF	Representasi IndoBERT dan Encoder-Decoder BiLSTM-CRF memperoleh F1-Score tertinggi sebesar 94,90%
3	*Fine-Tuning for Named Entity Recognition Using Part-of-Speech Tagging* (Masaya dkk., 2018)	IREX CRL yang terdiri dari 1.174 surat kabar Jepang "The Mainichi Shimbun"	POS2NER (penyempurnaan NER dengan penambahan POS) dan NER konvensional	POS2NER secara konsisten mencapai hasil yang lebih baik dibandingkan NER konvensional

Berikut adalah *resume* penelitian pada Tabel 1.

2.1.1 Penelitian Pertama (Andri dk., 2023)

Penelitian ini bertujuan untuk mengumpulkan informasi geografis dari teks klasifikasi yang berfokus pada ekstraksi lokasi serta menganalisis performa algoritma *Hybrid* BiLSTM dan CRF dengan penambahan POS *Tagger* menggunakan *Indonesian CRF-Tagger* pada kejadian lalu lintas. Penelitian ini menggunakan data berjumlah 615 tweets dataset yang diperoleh melalui pengumpulan data online secara otomatis oleh *web crawling* pada akun twitter lewatmana.com dalam rentang waktu 22 April sampai 18 Maret 2022.

Prosedur pelabelan menggunakan Label Studio. Selanjutnya tahap pos tagging sebagai prosedur pemberian tag dalam penelitian ini, penandaan POS dilakukan secara otomatis menggunakan sebuah CRF-Tagger Indonesia dimana sebelum POS Tagging dilakukan, teks harus sudah dilakukan tokenized. Model ekstraksi lokasi ini dikembangkan melalui lima komponen utama. Pertama, input data yang diterima oleh model ini dibatasi dengan panjang maksimal 25 kata. Selanjutnya, untuk merepresentasikan kata-kata dalam bentuk vektor, digunakan penyematan kata (word embedding), di mana setiap kata dipetakan ke dalam dimensi vektor berukuran 40. Setelah itu, model mengandalkan LSTM dua arah (BiLSTM) dengan penerapan dropout berulang sebesar 0.1 untuk mencegah overfitting dimana setiap kata dalam kalimat diberikan aktivasi menggunakan fungsi Relu. Hasil yang diperoleh adalah akurasi pada epoch 500 sebesar 91,01% grafik (loss dan akurasi) mengalami titik jenuh pada kurang dari 200 epoch. Oleh karena itu, uji coba kedua dilakukan dengan mengatur epoch menjadi 200. Model kemudian mencapai sedikit peningkatan sebesar 0,20% menjadi 91,21%. Serta dari penelitian juga ditemukan empat entitas yang membangun informasi kejadian lalu lintas dengan menggunakan teknik Latent Dirichlet Allocation (LDA), yaitu kondisi lalu lintas, lokasi, nama lokasi, nama jalan, infrastruktur, dan penyebab kejadian lalu lintas.

2.1.2 Penelitian Kedua (Siti Oryza Khairunnisa, 2021)

Penelitian ini bertujuan untuk melakukan anotasi ulang pada dataset NER Indonesia yang dianotasi untuk meningkatkan konsistensinya serta menawarkan dataset yang telah ditingkatkan untuk bahasa Indonesia, dengan memperbaiki dataset yang sudah ada sebelumnya. Melakukan pembelajaran lintas bahasa dari beberapa bahasa yaitu Bahasa Inggris, Spanyol, Belanda, dan Jerman, ke dalam

bahasa Indonesia dengan menggunakan BiLSTM dan CRF sebagai *encoder decoder* dengan membandingkan tiga model untuk percobaan dengan dua model multibahasa berbasis *transformer* BERT yaitu *Multilingual Bidirectional Encoder Representations from Transformers* (mBERT) dan *Cross-lingual Language Model - Robustly Optimized BERT Approach* (XLM-RoBERTa) dan dua BERT monolingual untuk bahasa Indonesia yaitu IndoBERT dan *Indonesian Language Evaluation Montage* (IndoLEM).

Dataset dibagi menjadi set uji yang sama dengan yang digunakan pada S&N pada tahun 2016 dan set pengembangan yang diambil secara acak dari set pelatihan. Penelitian ini menguji dua pendekatan transfer lintas bahasa tanpa pengawasan untuk NER bahasa Indonesia, yakni transfer sumber tunggal dengan bahasa Inggris dan transfer multi sumber. Data untuk pendekatan transfer sumber tunggal diambil dari dataset *Conference on Natural Language Learning 2003* (CoNLL-2003) dan korpora paralel bahasa Inggris Indonesia, sementara pendekatan multi sumber melibatkan beberapa bahasa sumber untuk memperluas cakupan transfer lintas bahasa. Hasil akhir penelitian ini menunjukan model dengan *input representations* IndoBERT dengan *encoder decoder* BiLSTM-CRF memperoleh skor tertinggi pada F1 skor yaitu sebesar 94.90%.

2.1.3 Penelitian Ketiga (Masaya dkk., 2018)

Penelitian ini menyelidiki dan mengembangkan metode yang lebih efektif dalam melakukan NER, khususnya untuk bahasa Jepang, dengan memanfaatkan teknik *fine tuning* yang mengintegrasikan informasi dari POS *tagging*. Penelitian berfokus meningkatkan kinerja NER melalui teknik *fine tuning* yang mengoptimalkan informasi dari POS *tagging*. Diharapkan pendekatan model NER akan mencapai akurasi dan efektivitas yang lebih tinggi, terutama dalam situasi di mana corpus NER yang tersedia terbatas.

Penelitian ini menggunakan data dari Information Retrieval and Extraction Exercise (IREX) Communications Research Laboratory (CRL) yang terdiri dari 1.174 artikel dari surat kabar Jepang ”*The Mainichi Shimbun*”. Data ini mengandung 19.262 tag entitas bernama yang mencakup berbagai kategori seperti *artifact*, *date*, dan *location*. Pembagian data dilakukan dengan rasio 3:1:1 untuk data pelatihan, set pengembangan, dan data uji. Dalam hal metode, penelitian

ini menggunakan format tagger *Inside, Outside, Beginning, End, Single* (IOBES) untuk anotasi dan POS *tagging* yang diambil dari *Universal Dictionary* (UniDic), dengan 21 jenis tag yang berbeda. Model yang digunakan dalam penelitian ini adalah *National Institute for Japanese Language and Linguistics Web Japanese Corpus* (NWJC2vec), sebuah model word2vec dengan dimensi 200 yang dilatih dari *National Institute for Japanese Language and Linguistics* (NINJAL) *Web Japanese Corpus*. Kinerja model dievaluasi dengan metrik *mikro-averaged precision (P), recall (R), dan F-measure (F)* untuk menilai seberapa baik model dapat mengenali entitas dalam teks. Hasil penelitian menunjukkan bahwa metode *Part-of-Speech* (POS) *tagging to Named Entity Recognition* (POS2NER) yang mengintegrasikan POS *tagging* selalu lebih baik dibandingkan dengan model NER tradisional dalam hal precision, recall, dan F-measure, kecuali pada beberapa kasus tertentu. Tabel hasil evaluasi yang menunjukkan kinerja model untuk tag seperti *artifact*, *date* dan *location* yang menunjukkan bahwa POS2NER lebih unggul pada *precision*, *recall* dan *F-measure*.

2.2 Artikel Berita Online

Artikel berita online merupakan salah satu media yang memiliki peran penting untuk memperluas penyampaian informasi melalui dunia maya (Ahsyar dan Afani, 2020). Informasi atau laporan yang dipublikasikan secara elektronik melalui internet. Berita ini dapat berupa tulisan, gambar, audio, atau video yang disampaikan melalui situs web berita, platform media sosial, aplikasi berita, atau email. Keberadaan berita online sangat vital dalam membentuk opini masyarakat dan meningkatkan kesadaran akan isu-isu penting di tingkat lokal maupun global.

2.3 Bahasa Indonesia

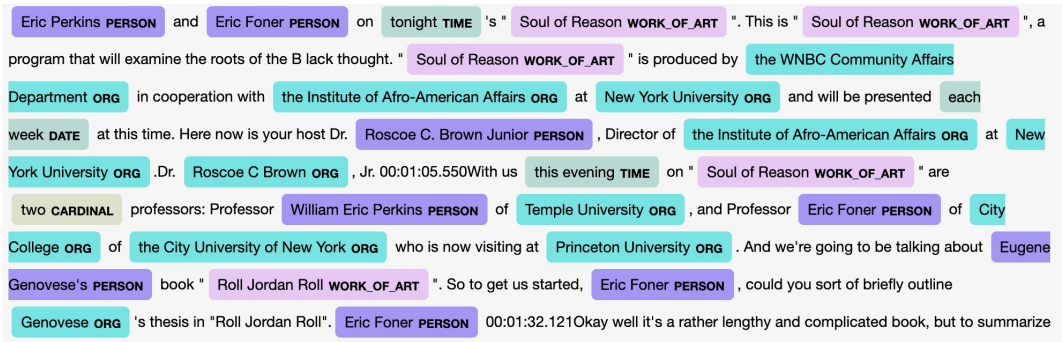
Bahasa indonesia terdiri dari dua kata yaitu ”Bahasa” dan ”Indonesia”, dimana setiap kata memiliki artinya tersendiri. Bahasa merupakan alat komunikasi yang penting bagi manusia, berfungsi sebagai media untuk menyampaikan pesan, ide, dan informasi antar individu. Sebagai bahasa resmi negara, bahasa Indonesia tidak hanya digunakan dalam interaksi sehari-hari, tetapi juga sebagai bahasa pengantar dalam pendidikan dan alat pengembangan pengetahuan, seni, serta teknologi (Maghfiroh, 2022).

2.4 Penyakit Menular

Penyakit menular memiliki dampak signifikan terhadap kesehatan masyarakat global. Karakteristik utama dari penyakit ini adalah keberadaan mikroorganisme patogen yang mampu berpindah dari satu inang ke inang lainnya, baik melalui transmisi langsung seperti kontak fisik atau droplet respiratori maupun transmisi tidak langsung melalui perantara seperti vektor biologis, makanan, air, udara, atau benda mati yang terkontaminasi. Patogen penyebab penyakit menular memiliki kemampuan untuk bertahan hidup di luar inang, bereplikasi dalam tubuh penderita, dan mengembangkan mekanisme transmisi yang kompleks, di mana efektivitas penularannya sangat dipengaruhi oleh interaksi antara tiga komponen utama yaitu agen penyebab penyakit, kerentanan inang, dan kondisi lingkungan yang mendukung proses penularan (Morens dan Fauci, 2020).

2.5 *Named Entity Recognition* (NER)

Named Entities (NEs) adalah kata benda yang mewakili kategori spesifik seperti nama orang, organisasi, dan lokasi. Konsep ini pertama kali diperkenalkan pada *Sixth Message Understanding Conference* (MUC-6), yang berfokus pada ekstraksi informasi dari sumber tidak terstruktur. Proses mengidentifikasi dan mengklasifikasikan informasi seperti nama individu, organisasi, lokasi, data temporal, serta informasi numerik dikenal sebagai NER (Bird, 2009). Metode NER merupakan salah satu jenis teknik dari NLP yang berguna untuk mengekstraksi entitas dari sebuah teks dan mengklasifikasikannya ke dalam kategori yang telah ditentukan sebelumnya (Roldos, 2020). Contoh penerapan NER disajikan pada Gambar 1 berikut:



Gambar 1. Contoh Penerapan NER
(Sumber: dandelion.eu)

Pelaksanaan tugas NER, entitas bernama dalam teks biasa ditandai menggunakan skema pelabelan BIO (Begin, Inside, dan Outside) dimana *Beginning of Named Entity* (B) digunakan untuk menandai awal dari sebuah entitas bernama, *Inside of Named Entity* (I) digunakan untuk menandai bagian tengah dari sebuah entitas bernama yang terdiri dari lebih dari satu kata, dan *Outside* (O) digunakan untuk menandai kata-kata yang bukan merupakan bagian dari entitas bernama dalam teks, seperti nama orang, organisasi, lokasi, dan yang lainnya (Lample dkk., 2016). Contoh teks: ”Barack Obama adalah Presiden Amerika Serikat” maka pelabelannya adalah I-per: Barack, B-per: Obama, other: adalah, B-loc: Amerika dan I-loc: Serikat. NER dapat dikategorikan menjadi beberapa jenis utama. Pertama, metode berbasis aturan yang menggunakan aturan secara manual. Kedua, metode *semi automatic labeling*, di mana model dilatih menggunakan dataset yang telah diberi label. Ketiga, metode *automatic labeling* (Konkol, 2015).

2.6 *Part Of Speech Tagging*

Part of Speech (POS) tagging atau secara singkat dapat ditulis sebagai POS *tagging* merupakan salah satu aspek penting dalam pemrosesan bahasa alami NLP yang bertujuan untuk memberikan label gramatikal pada setiap kata dalam sebuah teks. Tagging ini memberikan informasi mengenai definisi dan konteks kata, sehingga memungkinkan sistem untuk memahami struktur sintaksis kalimat. POS tagging tidak hanya digunakan untuk mengidentifikasi jenis kata seperti kata benda (*noun*), kata kerja (*verb*), atau kata sifat (*adjective*), tetapi juga berperan penting dalam berbagai aplikasi NLP seperti named entity recognition, analisis sentimen, dan penerjemahan mesin. Dengan menyematkan label gramatikal pada setiap token, sistem menjadi lebih mampu dalam mengenali pola dan struktur kalimat yang kompleks, terutama dalam bahasa alami seperti bahasa Indonesia. Keberhasilan POS tagging dalam meningkatkan kinerja model NLP telah dibuktikan dalam berbagai penelitian, salah satunya menunjukkan bahwa penggunaan POS sebagai fitur tambahan dapat memperbaiki performa dalam tugas ekstraksi informasi secara signifikan (Ratnaparkhi, 1996). Rincian POS *tagging* disajikan pada tabel 2 berikut:

Tabel 2. Kategori POS

No	POS	POS Name	Example
1	OP	Open Parenthesis	([
2	CP	Close Parenthesis	)]
3	GM	Slash	/
4	;	Semicolon	;
5	:	Colon	:
6	”	Quotation	”
7	.	Sentence Terminator	.
8	,	Comma	,
9	-	Dash	-
10	...	Elipsis	...
11	JJ	Adjective	Kaya, Manis
12	RB	Adverb	Sementara, Nanti
13	NN	Common Noun	Mobil
14	NNP	Proper Noun	Bekasi, Indonesia
:	:	:	:
35	FW	Foreign Words	Foreign, Word

Berdasarkan Tabel 2, korpus bahasa Indonesia menyediakan 35 kategori POS yang berbeda. Kategori ini mencakup berbagai jenis kata, seperti kata benda, kata kerja, kata sifat, dan kata keterangan. Setiap kategori memiliki fungsi spesifik dalam kalimat dan berkontribusi terhadap pemahaman makna keseluruhan teks. Misalnya, kata benda (NN) dan kata kerja (VBI) sering menjadi inti dari kalimat, sedangkan kata sifat (JJ) dan adverbia (RB) memberikan informasi tambahan yang memperjelas konteks (Wicaksono dkk., 2010).

2.7 *Deep Learning*

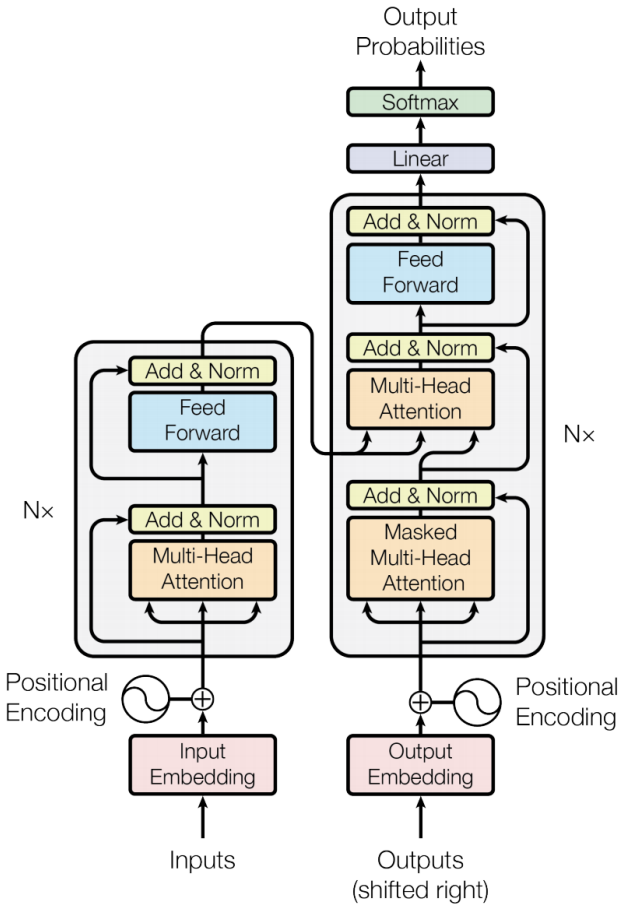
Deep Learning merupakan subset dari *machine learning* yang terinspirasi dari struktur jaringan saraf biologis manusia, yang memungkinkan komputer untuk belajar dari pengalaman dan memahami dunia sebagai hierarki konsep dan makna (LeCun, 2015). Teknologi ini menggunakan lapisan-lapisan jaringan saraf tiruan atau *artificial neural networks* (ANN) yang saling terhubung untuk memproses dan mentransformasikan data. Setiap lapisan dalam arsitektur *deep learning* memiliki

kemampuan untuk mengekstrak fitur dari input data secara otomatis, mulai dari fitur-fitur sederhana pada lapisan awal hingga fitur-fitur yang lebih kompleks pada lapisan yang lebih dalam.

Metode *deep learning* memiliki ciri khas yaitu penggunaannya memakai banyak lapisan atau biasa disebut sebagai ”*deep*” pada jaringan neural. Kemampuannya sangat berguna untuk memahami dan menganalisis data seperti pengenalan wajah, pengenalan suara, pelabelan text dan pemecahan masalah yang lebih kompleks lainnya. Beberapa algoritma *deep learning* antara lain BILSTM dan IndoBERT, dimana untuk algoritma IndoBERT sendiri masuk kedalam *transformer* karena termasuk dalam kategori *Deep Learning* yang dibangun di atas model BERT.

2.7.1 *Transformer*

Transformer adalah salah satu arsitektur model *deep learning* yang paling berpengaruh dalam bidang pemrosesan bahasa alami NLP dalam beberapa tahun terakhir. Model *transformer* diperkenalkan dalam makalah berjudul ”*Attention is All You Need*” (Vaswani dkk., 2017) dan telah mengubah lanskap NLP secara signifikan, mengungguli banyak pendekatan berbasis *Recurrent Neural Network* (RNN) atau *Convolutional Neural Network* (CNN) yang sebelumnya mendominasi. Pada dasarnya, model *transformer* mengandalkan mekanisme *self-attention*, yang memungkinkan model untuk fokus pada bagian yang relevan dari *input*, misalnya, kata-kata dalam kalimat tanpa mengandalkan urutan tertentu. Ini adalah perubahan signifikan dibandingkan dengan model RNN atau CNN yang sebelumnya sangat bergantung pada urutan data seperti teks dalam memproses informasi. Berikut penjelasan mengenai *transformer* dsajikan pada Gambar 2 berikut:



Gambar 2. Arsitektur *Transformer*
(Vaswani dkk., 2017)

Arsitektur *transformer* terdiri dari dua bagian utama yaitu *encoder* dan *decoder*. *Encoder* bertugas untuk memproses *input*, sementara *decoder* menghasilkan *output* berdasarkan representasi yang dibuat oleh *encoder*. Baik *encoder* maupun *decoder* terdiri dari lapisan *self-attention*. Misalkan sebuah input sekuens adalah $x = (x_1, x_2, \ldots, x_n)$ dan output sekuens adalah $y = (y_1, y_2, \ldots, y_m)$. *Encoder* memetakan input x ke representasi vektor $z = (z_1, z_2, \ldots, z_n)$, sedangkan *decoder* memetakan representasi z dan output sebelumnya $(y_1, y_2, \ldots, y_{t-1})$ ke output y_t saat ini. Pada *transformer* baik *encoder* dan *decoder* terdiri dari 6 lapisan identik, dengan masing-masing terdiri dari dua *sub*-lapisan yaitu *multi-head self-attention* dan jaringan *feed-forward* sederhana sedangkan untuk *encoder* terdapat tambahan lapisan *attention*.

Komponen krusial dalam arsitektur *Transformer* adalah mekanisme *Multi-Head Attention*. Mekanisme ini memungkinkan model untuk secara simultan

memperhatikan informasi dari berbagai posisi berbeda, dengan setiap *head attention* memfokuskan pada aspek *relasional* yang berbeda dalam data. Proses ini melibatkan transformasi linear input menjadi tiga representasi berbeda yaitu *Query, Key* dan *Value*, yang kemudian digunakan untuk menghitung skor *attention*. Setiap layer dalam *encoder* maupun *decoder* mengandung *sublayer Feed-Forward Network* yang terdiri dari dua *transformasi linear* dengan aktivasi *non-linear* di antaranya. Sublayer ini berperan dalam pemrosesan representasi yang dihasilkan oleh mekanisme *attention*. Untuk memfasilitasi pembelajaran yang efektif, setiap sublayer dilengkapi dengan *Layer Normalization* dan *Residual Connection*.

Transformer mengadopsi arsitektur *stack* yang terdiri dari *multiple layer* identik. Implementasi standar, baik *Encoder* maupun *Decoder* tersusun dari enam layer. Setiap layer *Encoder* mencakup *self-attention layer* dan *feed-forward network*, sementara layer *Decoder* menambahkan *cross-attention layer* yang memproses *output Encoder*. Keunggulan signifikan arsitektur Transformer terletak pada kemampuannya memproses seluruh sekuens secara paralel, mengeliminasi keterbatasan pemrosesan sekuensial yang menjadi karakteristik arsitektur RNN. Hal ini memungkinkan model untuk menangkap dependensi jarak jauh *(long-range dependencies)* dengan lebih efektif. Perkembangan terkini berbagai modifikasi dan optimasi terhadap arsitektur dasar *Transformer* terus diusulkan, mencakup efisiensi komputasional, skalabilitas, dan kemampuan pemrosesan sekuens yang lebih panjang.

2.7.2 Inisiasi Parameter *Hypertunning*

Hypertunning merupakan proses yang dilakukan sebelum pembelajaran model. Penyesuaian ini berguna untuk memaksimalkan kinerja model saat akan melakukan proses validasi. Beberapa proses *Hypertunning* yang dilakukan seperti penambahan *epoch, batch size, optimizer* dan fungsi aktivasi. Parameter yang digunakan pada proses *hypertunning* yaitu:

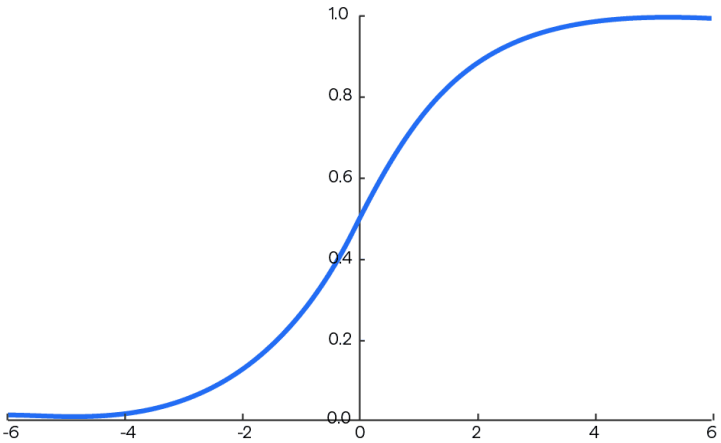
1. *Epoch* merupakan iterasi dimana ketika semua dataset telah melalui proses *training* pada model maka akan dikembalikan ke awal putaran pertama (Brownlee, 2016)
2. *Batch size* digunakan untuk pembagian pada *epoch* menjadi bagian-bagian kecil untuk mempercepat proses *hypertunning*. Hal ini dilakukan karena proses *epoch* membutuhkan waktu yang lama. Ukuran batch ideal berkisar 64-128, yang memberikan keseimbangan antara kecepatan komputasi dan kualitas

pemelajaran model, dengan memperhatikan karakteristik spesifik dataset dan arsitektur jaringan (Hwang dkk., 2024).

3. *Optimizer* memliki tujuan utama untuk meminimalkan atau memaksimalkan fungsi *loss* dengan cara menemukan nilai parameter yang optimal sehingga model dapat membuat prediksi yang lebih akurat. Beberapa algoritma populer adalah SGD, Adam, RMSprop, Adagrad, dan Momentum (Mahajaya dkk., 2024).
4. *Activation function* adalah komponen penting dalam jaringan saraf (*neural network*) yang menentukan keluaran dari neuron berdasarkan input yang diterima. Fungsi aktivasi memungkinkan model untuk memperkenalkan *non-linearitas* dalam jaringan, yang memungkinkan jaringan untuk mempelajari dan menangkap pola yang lebih kompleks dan lebih akurat dalam data. Beberapa fungsi aktivasi yang sering digunakan antara lain Sigmoid, Tanh, ReLU, Softmax, dan ELU (Wibawa, 2017).

2.7.3 Fungsi Aktivasi *Softmax*

Fungsi aktivasi *softmax* adalah fungsi aktivasi yang biasa digunakan pada lapisan *output* dari jaringan saraf tiruan, khususnya untuk kasus klasifikasi multi-kelas (Ilahiyah dan Nilogiri, 2018). Fungsi *softmax* memainkan peran krusial sebagai fungsi aktivasi pada lapisan *output*. Pada implementasinya, *softmax* mengambil vektor nilai *real* dari lapisan sebelumnya dan menormalisasikannya menjadi probabilitas yang dapat diinterpretasikan. Grafik fungsi *softmax* disajikan pada Gambar 3 berikut:



Gambar 3. Grafik Fungsi *Softmax*
(BotPenguin.com)

Fungsi yang diberikan pada Persamaan (2.1) (Kouretas dan Paliouras, 2020):

$$f_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}} \tag{2.1}$$

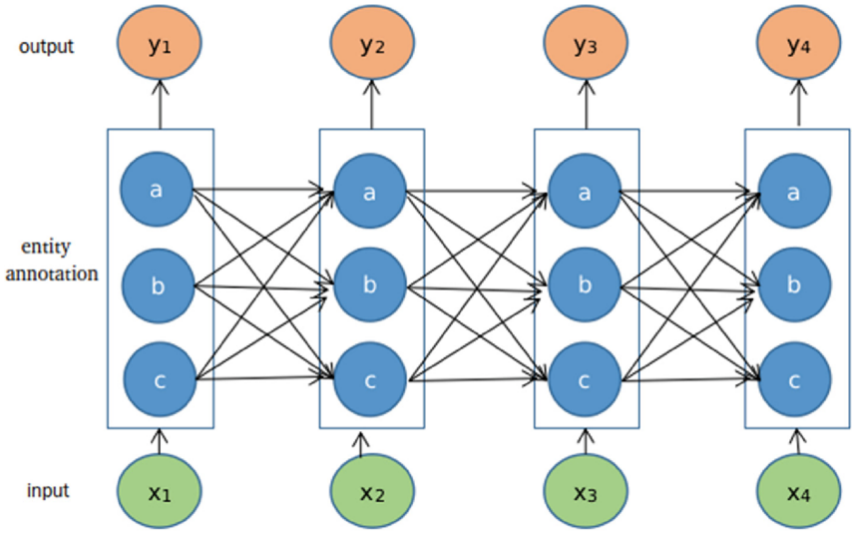
dengan:

- e adalah bilangan Euler (sekitar 2,71828),
- z_j adalah nilai input untuk kelas j,
- n adalah jumlah total kelas,
- $\sum_{k=1}^{n} e^{z_k}$ adalah penjumlahan dari e^{z_k} untuk setiap kelas k, dari 1 hingga n.

Softmax mengubah output skor mentah dari model ke dalam bentuk probabilitas yang bisa digunakan untuk memprediksi kelas suatu input. Probabilitas yang dihasilkan dari fungsi Persamaan (2.1) selalu berada di antara 0 dan 1 dan jumlah totalnya adalah 1.

2.7.4 *Conditional Random Field* (CRF)

Conditional Random Field (CRF) adalah model probabilistik yang digunakan untuk prediksi terstruktur, sering diterapkan dalam pemrosesan bahasa alami atau *Natural Language Processing* (NLP) dan pengenalan pola. Model CRF memungkinkan pengambilan keputusan yang mempertimbangkan konteks dari data input, berbeda dengan model klasifikasi tradisional yang hanya memprediksi label untuk satu sampel tanpa mempertimbangkan sampel tetangga (Khan dkk., 2021). Model CRF memungkinkan urutan output diprediksi dengan mempertimbangkan hubungan antar elemen dalam urutan tersebut, yang direpresentasikan dalam bentuk graf (Fu, 2017). Model ini sangat efektif untuk tugas-tugas seperti POS *tagging* dan NER. Arsitektur CRF disajikan pada Gambar 4 berikut:



Gambar 4. Arsitektur *Conditional Random Field (* (CRF)

Rumus yang diberikan untuk model *linear chain Conditional Random Field* (CRF) dijelaskan pada Persamaan (2.2) berikut:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right) \tag{2.2}$$

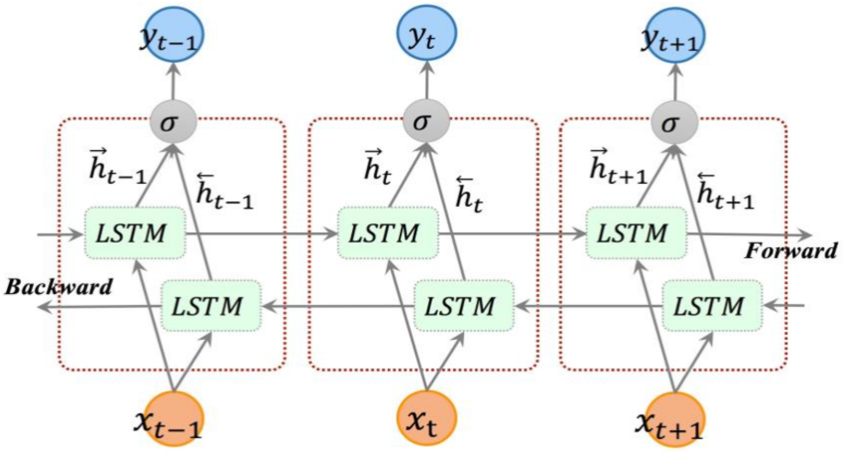
dengan:

- $p(y|x)$ adalah probabilitas dari urutan label y diberikan urutan input x.
- $Z(x)$ adalah fungsi normalisasi yang memastikan total probabilitas sama dengan satu. Fungsi ini dihitung dengan menjumlahkan eksponensial dari semua kemungkinan urutan label, sehingga menjamin bahwa probabilitas yang dihasilkan valid (yaitu, totalnya 1).
- λ_k adalah bobot untuk fitur ke-k, yang menunjukkan seberapa besar kontribusi fitur tersebut terhadap prediksi akhir.
- $f_k(y_t, y_{t-1}, x_t)$ adalah fungsi fitur yang mengukur kekuatan hubungan antara entitas dalam urutan label dan input. Fungsi ini berfungsi untuk mengekstrak informasi dari data dan mengidentifikasi pola yang relevan dalam konteks NER.

Model *linear chain* CRF sangat berguna dalam aplikasi di mana urutan data menjadi penting. Model ini dapat menangkap ketergantungan antar label-label, memungkinkan prediksi yang lebih akurat dan tepat berdasarkan konteks sekitarnya (Sutton dan McCallum, 2012).

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

Bidirectional Long Short-Term Memory merupakan arsitektur jaringan saraf yang sangat efektif dalam pemrosesan bahasa alami, khususnya dalam tugas-tugas seperti NER dan POS *Tagging*. Model BiLSTM mengatasi keterbatasan dari model LSTM tradisional . Model ini memiliki dua jaringan LSTM. Jaringan LSTM pertama bertugas untuk memproses urutan data dari arah depan (*forward*), sementara jaringan LSTM kedua berfungsi untuk memproses urutan data dari arah sebaliknya (*backward*). Output dari kedua jaringan LSTM, baik yang maju maupun mundur, digabungkan pada setiap langkah waktu. Dua lapisan yang berlawanan arah ini membuat model mampu mempelajari informasi dari masa lalu serta informasi dari masa depan untuk setiap urutan input. Arsitektur BiLSTM disajikan pada Gambar 5 berikut:



Gambar 5. Arsitektur BiLSTM

Arsitektur BiLSTM melakukan proses pada *forward* dan *backward* dimana gabungan *forward* dan *backward* dapat ditulis sebagai (Hamed dan Zapirain, 2020):

• *Forward* LSTM dituliskan dalam Persamaan (2.3) berikut:

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}}) \tag{2.3}$$

dengan:

- x_t adalah input pada waktu t,
- $\overrightarrow{h_{t-1}}$ adalah *hidden state* pada waktu $t-1$.

• *Backward* LSTM dituliskan dalam Persamaan (2.4) berikut:

$$\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}}) \tag{2.4}$$

dengan:

- x_t adalah input pada waktu t,
- $\overleftarrow{h_{t+1}}$ adalah *hidden state* pada waktu $t + 1$.

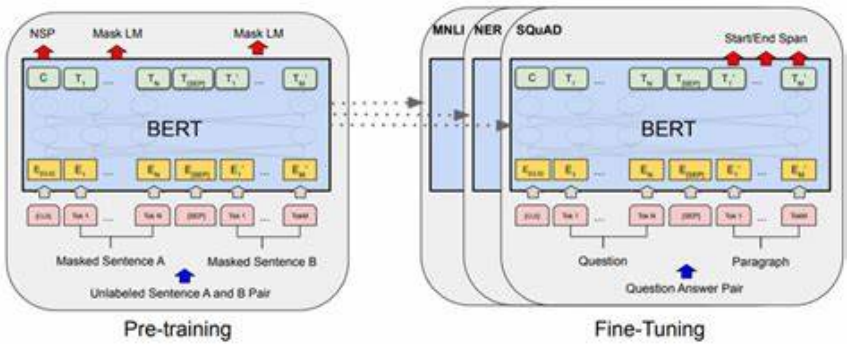
• *Output* Akhir Bi-LSTM dituliskan dalam Persamaan (2.5) berikut::

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{2.5}$$

Output akhir adalah gabungan dari *hidden state* maju ($\overrightarrow{h_t}$) dan *hidden state* mundur ($\overleftarrow{h_t}$).

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

Indonesian Bidirectional *Encoder Representations from Transformers* merupakan *pre-trained language* model yang dikembangkan khusus untuk Bahasa Indonesia, menggunakan arsitektur yang sama pada BERT (Wilie dkk., 2020). Model ini dilatih menggunakan dataset berbahasa Indonesia untuk menghasilkan pemahaman bahasa yang baik untuk konteks lokal. Model ini memiliki beberapa karakteristik penting yang membedakannya dari BERT original, dimana IndoBERT menggunakan vocabulary yang dikhususkan untuk Bahasa Indonesia dengan jumlah 31.944 token (Koto dkk., 2020). Berikut artsitektur BERT disajikan pada Gambar 6 berikut:



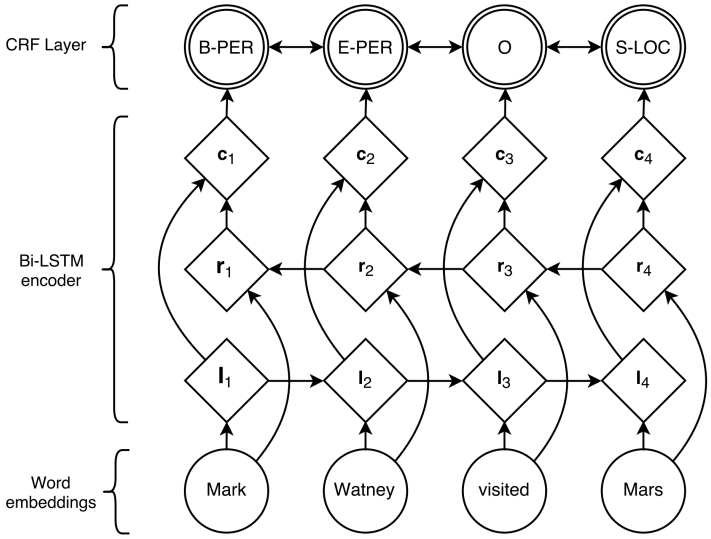
Gambar 6. Arsitektur BERT
(Sumber: velog.io)

Pelatihan IndoBERT dilakukan dengan menggunakan dataset besar yang mencakup berbagai domain dalam bahasa Indonesia, seperti korpus berita, media sosial, dan literatur online. Dengan demikian, IndoBERT tidak hanya mengandalkan pengenalan kata dasar, tetapi juga dapat menangkap variabilitas bahasa yang sering muncul dalam percakapan sehari-hari atau penggunaan bahasa tidak formal. Pada intinya, IndoBERT adalah versi lokal dari BERT yang dilatih untuk memahami dan memproses teks dalam bahasa Indonesia secara lebih efektif.

Salah satu tantangan utama dalam NLP untuk bahasa Indonesia adalah pengelolaan variabilitas kata dan struktur kalimat yang tidak selalu mengikuti pola yang jelas seperti dalam bahasa Inggris. Sebagai contoh, dalam bahasa Indonesia terdapat banyak kata majemuk dan penggunaan kata yang sering dipengaruhi oleh konteks sosial dan budaya. IndoBERT, dengan menggunakan teknik *subword tokenization*, mampu menangani kata-kata yang sebelumnya tidak dikenal atau memiliki variasi bentuk yang luas. Dengan pendekatan ini, IndoBERT bisa mengurangi masalah yang sering muncul pada model berbasis kata (*word-based models*), yang terkadang kesulitan dalam menangani kata-kata baru.

2.7.7 *Hybrid* BiLSTM-CRF

Hasil dari kedua arah kemudian dikombinasikan untuk meningkatkan kemampuan model dalam memodelkan dependensi sekuens secara lebih lengkap. Sedangkan model BiLSTM dengan penambahan layer CRF ditampilkan pada Gambar 7 berikut:



Gambar 7. Arsitektur BiLSTM-CRF
(researchgate.net)

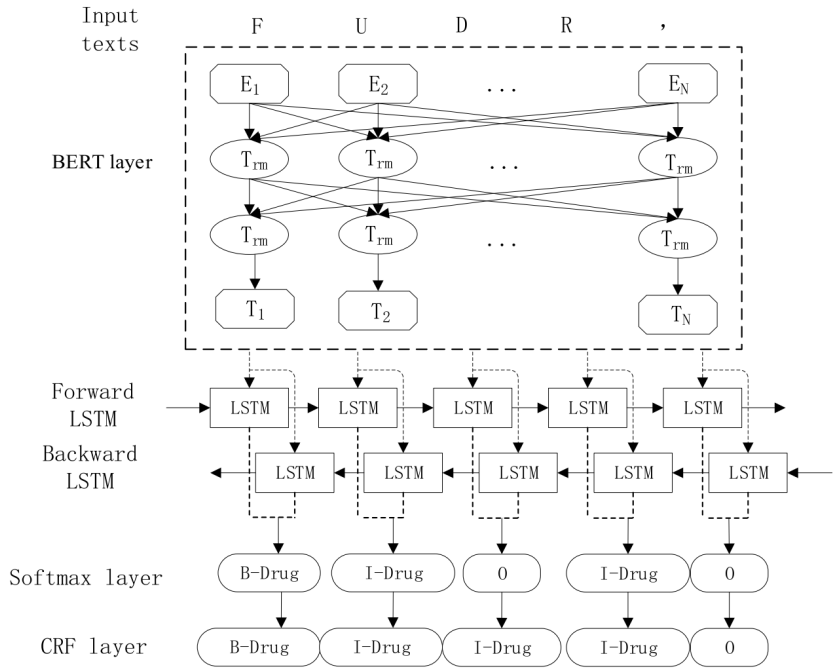
Model BiLSTM-CRF dalam NER menggabungkan kelebihan dari BiLSTM dalam memahami konteks sekuensial dua arah dan CRF dalam memastikan konsistensi antar label dalam sekuens. Gabungan ini menjadikannya sangat efektif dalam mengenali entitas dalam teks dengan akurasi yang tinggi, terutama ketika entitas tersebut memiliki struktur yang kompleks atau konteks yang ambigu.

2.7.8 *Hybrid* IndoBERT-BiLSTM-CRF

Penggunaan IndoBERT dalam model ini bertujuan untuk memperoleh representasi kata yang kontekstual, sesuai dengan struktur dan gaya bahasa Indonesia. Berbeda dengan model embedding statis seperti Word2Vec atau GloVe, IndoBERT mampu memahami makna kata berdasarkan konteks kalimat secara menyeluruh. Hal ini sangat penting dalam tugas NER, karena satu kata dapat memiliki makna berbeda tergantung pada posisinya dalam kalimat.

Setelah embedding dihasilkan oleh IndoBERT, representasi kata tersebut diproses lebih lanjut oleh lapisan BiLSTM. Lapisan ini bekerja secara dua arah, yaitu dari kiri ke kanan dan dari kanan ke kiri, untuk menangkap informasi kontekstual dari seluruh urutan kata. Dengan demikian, BiLSTM mampu memperkaya pemahaman model terhadap hubungan antar kata dalam kalimat secara lebih mendalam, yang sangat penting untuk mendeteksi entitas secara akurat.

Pada tahap akhir, hasil keluaran dari BiLSTM diteruskan ke lapisan CRF. Lapisan ini digunakan untuk memastikan bahwa urutan label yang dihasilkan sesuai dengan struktur yang valid dalam tugas NER. Dengan mempertimbangkan hubungan antar label, CRF mampu meningkatkan akurasi prediksi dan mengurangi kemungkinan kesalahan, seperti munculnya label “I-LOC” tanpa didahului oleh “B-LOC”. Kombinasi ketiga komponen ini membentuk arsitektur hybrid yang kuat dan efektif dalam menyelesaikan permasalahan Named Entity Recognition dalam bahasa Indonesia. Artsitektur penggunaan BERT-BiLSTM-CRF ditampilkan pada Gambar 8 dibawah ini.



Gambar 8. Arsitektur BERT-BiLSTM-CRF
(Sumber: Gao dkk., 2021)

Model *hybrid* IndoBERT-BiLSTM-CRF menggabungkan kekuatan pemahaman konteks dari IndoBERT, pemrosesan urutan dua arah dari BiLSTM, dan kemampuan CRF dalam menghasilkan urutan label yang konsisten. Integrasi ini memungkinkan model mengenali entitas secara lebih akurat dan sesuai dengan struktur bahasa Indonesia.

2.8 Evaluasi Kinerja Model

Matriks evaluasi digunakan untuk mengukur performa dari model yang telah dibangun. Tujuannya adalah untuk mengetahui sejauh mana model dapat menghasilkan prediksi yang akurat. Salah satu bentuk evaluasi yang paling umum adalah confusion matrix. Confusion matrix memberikan gambaran jumlah prediksi yang benar dan salah dari masing-masing kelas.

2.8.1 *Confusion Matrix Kelas Biner*

Confusion matrix merupakan pembelajaran yang berisi informasi mengenai klasifikasi aktual dan prediksi. *Confusion matrix* terdiri dari empat komponen yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), serta *False Negative* (FN), yang berarti sistem memprediksi negatif tetapi hasilnya salah (Rabbani dkk., 2023). Istilah tersebut biasa dirangkum sebagai suatu matrix yang disebut confusion matrix sebagai ditunjukan pada Tabel 3 sebagai berikut:
Tabel 3. *Confusion matrix*

	Aktual Positif	**Aktual Negatif**
Prediksi Positif	*True Positive* (TP)	*False Positive* (FP)
Prediksi Negatif	*False Negative* (FN)	*True Negative* (TN)

- *True Positif* (TP), jumlah data aktual positif dan model prediksinya juga positif.
- *True Negative* (TN), jumlah data aktual negatif dan model prediksinya juga negatif.
- *False Positive* (FP), jumlah data aktual negatif dan model prediksinya positif.
- *False Negative* (FN), jumlah data aktual positif dan model prediksinya negatif.

Confusion matrix digunakan untuk mengevaluasi performa model klasifikasi dengan lebih rinci dibandingkan hanya mengandalkan akurasi saja. Setiap komponen dalam confusion matrix memberikan informasi penting tentang jenis kesalahan yang dilakukan model, dari confusion matrix, berbagai metrik evaluasi dapat dihitung untuk memahami kinerja model secara lebih mendalam.

2.8.2 *Confusion Matrix Multi Class*

Klasifikasi *multiclass* atau banyak kelas, *confusion matrix* memiliki struktur yang lebih kompleks untuk mengevaluasi performa model. Matriks ini disusun dengan baris yang merepresentasikan kelas sebenarnya (*actual class*) dan kolom yang merepresentasikan kelas hasil prediksi (*predicted class*). Nilai-nilai pada diagonal utama matriks dari pojok kiri atas ke pojok kanan bawah menunjukkan jumlah prediksi yang tepat untuk masing-masing kelas. Sedangkan nilai-nilai di luar diagonal utama mengindikasikan kesalahan klasifikasi, dimana model salah memprediksi suatu kelas sebagai kelas yang lain. Terlihat pada Tabel 4 dibawah yang menunjukan *confusion matrix* untuk *multiclass*.

Tabel 4. *Confusion Matrix* Untuk *Multi Class*

	A	**B**	**C**
A	True (AA)	False (AB)	False (AC)
B	False (BA)	True (BB)	False (BC)
C	False (CA)	False (CB)	True (CC)

Studi kali ini akan digunakan empat jenis matriks evaluasi yaitu sebagai berikut (Kurniasari dkk., 2024):

1. *Accuracy*: Persamaan akurasi dinyatakan pada persamaan (2.6) sebagai berikut:

$$Acc = \frac{\sum_{i=1}^{N} TP(L_i)}{\sum_{i=1}^{N} \sum_{j=1}^{N} L_{i,j}} \tag{2.6}$$

   Akurasi digunakan untuk mengukur proporsi dari prediksi yang benar yaitu total keseluruhan seberapa sering model benar mengklasifikasi.

2. *Precision*: Persamaan presisi dinyatakan pada persamaan (2.7) sebagai berikut:

$$P(L_i) = \frac{TP(L_i)}{TP(L_i) + FP(L_i)} \tag{2.7}$$

   Presisi digunakan untuk mengukur seberapa akurat model dalam mengidentifikasi kelas tertentu dari semua prediksi positif yang dibuat oleh model.

3. *Recall* (Sensitivitas atau *True Positive Rate*): Persamaan recall dinyatakan pada

persamaan (2.8) sebagai berikut:

$$R(L_i) = \frac{TP(L_i)}{TP(L_i) + FN(L_i)} \tag{2.8}$$

Recall merupakan matriks yang digunakan untuk menghindari *missed detection* atau gagal mendeteksi kasus positif yang sebenarnya.

4. *F1-score:* Persamaan *F1-score* dinyatakan pada persamaan (2.9) sebagai berikut:

$$F1(L_i) = \frac{2 \times P(L_i) \times R(L_i)}{P(L_i) + R(L_i)} \tag{2.9}$$

F1-score adalah rata-rata harmonik dari *precision* dan *recall*, memberikan gambaran tentang kinerja model secara keseluruhan dalam mengklasifikasikan kelas tertentu.

BAB III

METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian dilaksanakan selama periode semester genap tahun ajaran 2024/2025. Penelitian ini pertama kali dilaksanakan di Pusat Riset Informatika, Badan Riset dan Inovasi Nasional (BRIN) yang berlokasi di Kota Bandung, Provinsi Jawa Barat. Pusat Riset Informatika BRIN merupakan salah satu lembaga penelitian nasional yang memiliki fokus pada pengembangan ilmu pengetahuan dan teknologi di bidang informatika. Selanjutnya, kegiatan penelitian dilanjutkan di lingkungan akademik, yaitu di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA), Universitas Lampung. Jurusan tersebut berada di Jalan Prof. Dr. Soemantri Brojonegoro No. 1, Gedung Meneng, Kota Bandar Lampung, Provinsi Lampung.

3.2 Data dan Alat

3.2.1 Data

Dataset yang digunakan dalam penelitian ini diberi nama ”riim4-antaradetik04”, sebuah koleksi data yang dikembangkan dan dikelola oleh BRIN. Dataset ini berisi judul dari artikel berita online terkait penyakit menular seperti cacar monyet, cacar air, tuberkolosis, demam berdarah dan rabies yang berasal dari dua portal berita Indonesia yaitu Detik dan Antara. Periode pengumpulan data dilakukan selama lima bulan, dimulai dari Januari 2020 hingga Mei 2020. Total 1592 baris data dengan 4 kolom yang merepresentasikan entitas-entitas dalam teks berita. Dataset tersebut direpresentasikan pada Tabel 5 berikut:

Tabel 5. Dataset riim4-antaradetik04

mongo id	portal	title	published at iso
655c312976 d33e0394 9e2124	Antara	Dinkes Mataram minta warga terapkan PHBS antisipasi cacar monyet	16/11/2023 19:03
6559339b76 d33e0394 9e0eb5	Antara	Sudah 1.500 lebih warga Bangladesh meninggal karena demam berdarah	16/11/2023 16:40
655c312e76 d33e0394 9e2125	Antara	Dinkes Biak tingkatkan kepaspadaan dini cegah penyakit cacar monyet	15/11/2023 18:45
65551068926 a9ddac 21d1006	Antara	Pemprov DKI siapkan strategi jangkau seluruh hewan penular rabies	15/11/2023 13:31
655c313476 d33e0394 9e2126	Antara	Hanya cacar air Dinkes pastikan Lombok Tengah bebas cacar monyet	15/11/2023 12:22

Adapun contoh dari salah satu portal berita online, yaitu portal berita Antara yang mengangkat informasi mengenai penyebaran cacar monyet di Indonesia dalam artikel berjudul ”Gubernur: RS di Jabar sudah siap hadapi kasus cacar monyet”. Artikel ini menginformasikan perkembangan terkini mengenai penyebaran penyakit cacar monyet yang telah terdeteksi di wilayah Jabar dan RS telah siap terhadap kasus tersebut. Berita ini disampaikan oleh penjabat gubernur Jawa Barat Bey Triadi Machmudin di Gedung Sate Bandung sebagai respons terhadap wabah yang mulai mencuat dan menciptakan perhatian di masyarakat. Contoh berita tersebut disajikan pada Gambar 9 berikut:

Gambar 9. Contoh Berita Online Portal Antara

Label studio memungkinkan untuk memilih kategori atau label yang sesuai dengan objek yang terdeteksi dalam gambar atau teks, pada topik ini fokus utama pelabelan adalah lokasi, orang dan organisasi dalam teks.

3.2.1 Alat

Peralatan yang digunakan dalam menunjang penelitian ini antara lain:

a. Perangkat Keras (*Hardware*)

Penelitian ini menggunakan laptop Acer Aspire ES1-420-518B dengan tipe 64-bit *operating system*, x64-*based processor*. Spesifikasi hardware perangkat tersebut adalah sebagai berikut:

Processor : AMD A4-5000 APU *with Radeon (TM) HD Graphics*
Memori : SSD 450 GB
RAM : 4 GB

b. Perangkat Lunak (*Software*)

Penelitian ini memanfaatkan perangkat lunak untuk melakukan pelabelan data menggunakan metode *Named Entity Recognition* (NER) dengan penambahan *POS tagging* dengan model BiLSTM, BiLSTM-CRF, dan IndoBERT-BiLSTM-CRF. Perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut:

1. *Google colab 1.0.0*

Google colab yaitu alat yang disediakan oleh Google untuk memudahkan

penggunaan *Colaboratory*, yang merupakan layanan *notebook* gratis berbasis web. Paket *google colab* versi 1.0.0 memiliki beberapa dependensi penting seperti *google-auth, ipykernel, ipyparallel, ipython, notebook, pandas, portpicker, requests, dan tornado*.

2. *Python 3.10.12*

   Python merupakan versi terbaru dari bahasa pemrograman *python* yang dirilis pada tahun 2023. kompatibel pada sistem operasi Windows. Versi ini menambahkan beberapa fitur baru dan perbaikan bug yang terdapat pada versi sebelumnya untuk meningkatkan kinerja program*python*.

3. *Library Pandas 2.2.2*

   Pandas tersedia di google colab yang merupakan versi internal atau pra-rilis yang belum dirilis secara luas. *Library* ini digunakan dalam analisis data dan manipulasi data terstruktur dalam mengolah dan menganalisis data dalam bentuk tabel.

4. *Library Sklearn-crfsuite 0.5.0*

   Sklearn-crfsuite berfungsi sebagai *wrapper* untuk *CRFsuite*, yaitu implementasi cepat dari *Conditional Random Fields* (CRFs). Library ini menyediakan antarmuka yang mirip dengan *scikit-learn*, sehingga memudahkan dalam melatih dan menerapkan model CRF untuk *sequence labeling*. *Sklearn-crfsuite* kompatibel dengan *scikit-learn*, memungkinkan teknik seperti *grid search* untuk mencari *hyperparameter* terbaik.

5. *Library Natural Language Toolkit 3.8.1*

   Natural Language Toolkit digunakan untuk pengolahan bahasa alami (NLP). *Library* ini menyediakan alat dan sumber daya untuk analisis teks, seperti *tokenisasi, stemming, lemmatization, parsing*, dan pengenalan entitas bernama dengan mencakup kumpulan korpus teks besar.

6. *Library Matplotlib 3.8.0*

   Matplotlib digunakan untuk membuat visualisasi data dalam berbagai format, seperti grafik garis, batang, *histogram*, dan *scatter plot*. Merupakan versi terbaru dari *matplotlib* pada bahasa pemprograman *python*.

7. *Library TensorFlow 2.17.0*

   Tenserflow merupakan sebuah library *open-source* yang digunakan untuk *machine learning* (ML) dan *deep learning*. *TensorFlow* menyediakan berbagai alat untuk membangun dan melatih model-model kecerdasan buatan untuk *supervised learning* dan *unsupervised learning*.

8. *Library Numpy 1.26.4*

   NumPy adalah salah satu pustaka *python* yang paling umum digunakan untuk

komputasi numerik dan manipulasi *array* multidimensi serta funsi-fungsi matematika yang dapat dioperasikan pada *array*.

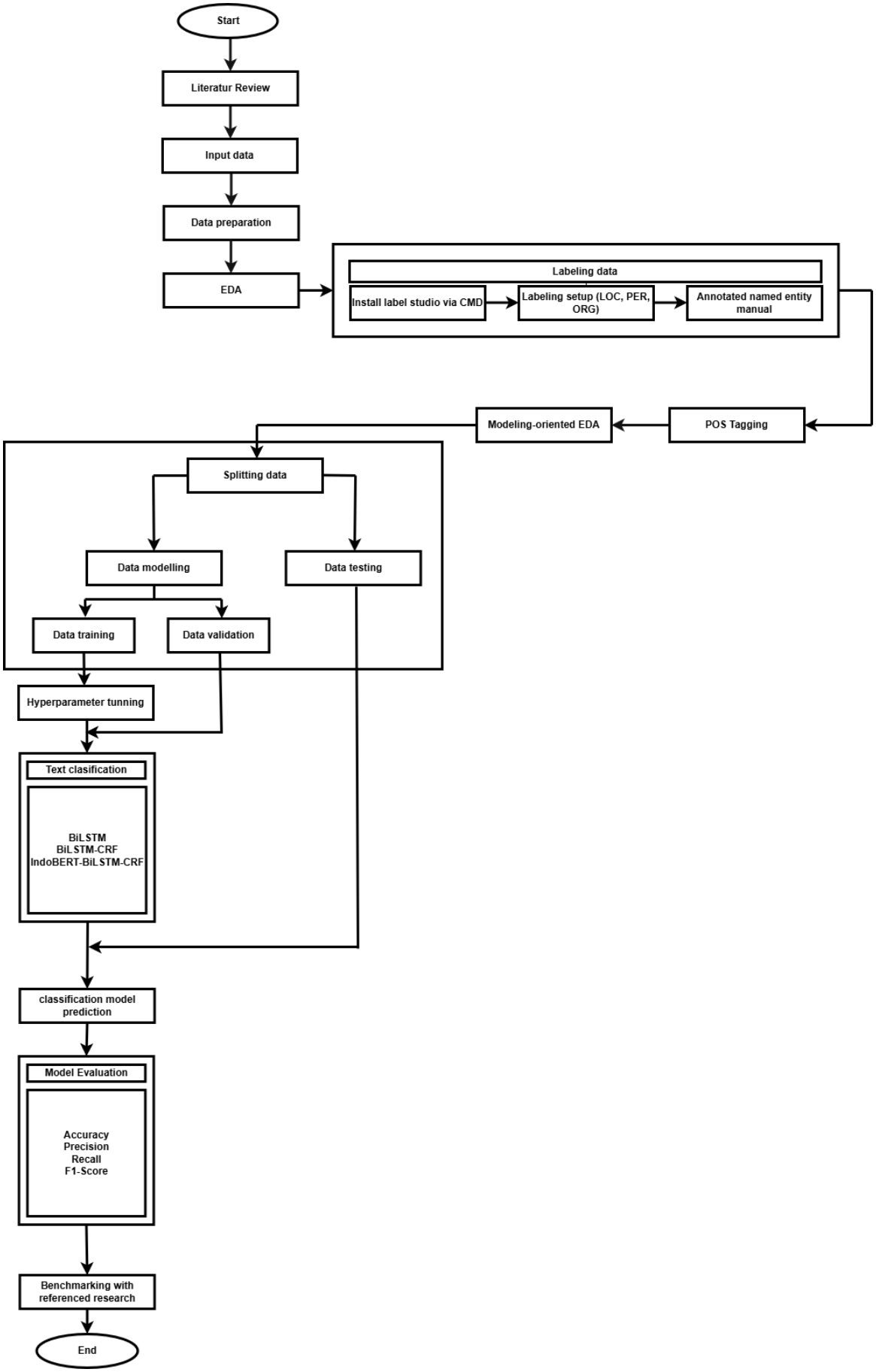
3.3 Alur Kerja Penelitian

Alur kerja penelitian ini mencakup serangkaian tahapan yang bertujuan untuk menentukan hasil klasifikasi pelabelan entitas menggunakan metode NER. Proses ini dilakukan dengan memanfaatkan algoritma *deep learning* berbasis arsitektur *transformer*, serta didukung oleh penggunaan platform *Google Colaboratory* untuk pengolahan dan analisis data.

1. Tahap studi literatur, akan dilakukan penelusuran dan perujukan terhadap berbagai sumber informasi yang tepercaya dan relevan guna mendukung landasan teori serta memperkuat kerangka penelitian.
2. Data dimasukkan ke dalam *Google Colab* melalui integrasi dengan *Google Drive*, sehingga memungkinkan akses dan pengolahan data secara langsung dari penyimpanan *cloud*.
3. Setelah data dipersiapkan, dilakukan *Exploratory Data Analysis* (EDA) yang mencakup pemeriksaan informasi umum dataset, analisis statistik deskriptif, identifikasi data hilang dan duplikat, serta visualisasi distribusi data berdasarkan portal.
4. Langkah selanjutnya adalah melakukan pelabelan manual menggunakan Label Studio, yang terdiri dari tiga tahap. Pertama, menginstal Label Studio melalui *Command Prompt* (CMD). Kedua, melakukan pengaturan proyek pelabelan dengan fokus pada entitas lokasi, orang, dan organisasi dan ketiga, melaksanakan proses pelabelan manual terhadap data yang telah disiapkan.
5. Setelah proses pelabelan data selesai, tahap selanjutnya adalah melakukan *Part of Speech* (POS) tagging secara otomatis menggunakan CRF-Tagger khusus untuk bahasa Indonesia yaitu ”all_indo_man_tag_corpus_model.crf.tagger”.
6. Tahap selanjutnya adalah *Modeling-Oriented Exploratory Data Analysis* (EDA), yang dilakukan setelah tahapan prapemrosesan data, pelabelan entitas, dan POS tagging. Tahapan ini mencakup analisis distribusi label entitas, distribusi POS, serta distribusi panjang kalimat untuk memahami karakteristik data sebelum memasuki tahap pemodelan.
7. Setelah data dipersiapkan, langkah selanjutnya adalah melakukan pembagian data (*data splitting*) menjadi dua bagian, yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model NER, sedangkan data pengujian

digunakan untuk mengevaluasi kinerja model Proporsi pembagian data dalam penelitian ini adalah 80% untuk pelatihan dan 20% untuk pengujian.

8. Setelah proses pembagian awal, data pelatihan (*data modeling*) kemudian dibagi kembali menjadi data pelatihan dan data validasi. Data pelatihan digunakan untuk melatih model, sementara data validasi berfungsi untuk mengatur parameter dan mencegah overfitting.
9. Setelah pembagian data selesai, tahap selanjutnya adalah melakukan *hyperparameter tuning* untuk memperoleh konfigurasi terbaik dalam proses pelatihan model. Penelitian ini mengimplementasikan dan membandingkan tiga arsitektur model, yaitu BiLSTM, BiLSTM-CRF, dan IndoBERT-BiLSTM-CRF.
10. Evaluasi model dilakukan menggunakan empat metrik utama, yaitu akurasi, *presisi*, *recall*, dan F1-*score*. Setelah evaluasi dilakukan, tahap selanjutnya adalah *benchmarking* untuk membandingkan performa antar model.



Gambar 10. Alur Kerja Penelitian

BAB V

KESIMPULAN DAN SARAN

Bab ini bertujuan untuk menyajikan kesimpulan dari hasil penelitian yang telah dilakukan dan saran mengenai penelitian ini. Adapun jabarannya sebagai berikut.

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah didapatkan diambil kesimpulan sebagai berikut.

1. Penerapan metode *Named Entity Recognition* (NER) berhasil digunakan untuk mengekstraksi informasi penting dari berita online tentang penyakit menular. Metode ini mampu mengenali dan memberi label pada entitas seperti lokasi, organisasi, serta tokoh atau pihak yang terlibat, sehingga membantu proses ekstraksi informasi menjadi lebih efisien dan sistematis.
2. Model yang dibangun dengan inputan representasi IndoBERT dan *encoder decoder* BiLSTM-CRF yang digunakan untuk mengklasifikasi berita online terkait penyakit menular dengan pembagian data 80% data *modeling* dan 20% data testing, kemudian data dibagi lagi menjadi data training sebesar 80% dan data validasi 20% menghasilkan model klasifikasi dan hasil terbaik berdasarkan akurasi, *precision*, *recall* dan *f1-score* yang diperoleh.

5.2 Saran

Saran yang dapat diberikan kepada pembaca dan peneliti selanjutnya adalah sebagai berikut.

1. Penelitian ini dapat dikembangkan lebih lanjut dengan melakukan modifikasi atau peningkatan terhadap arsitektur model, termasuk eksplorasi parameter yang lebih kompleks atau penggunaan teknik lanjutan seperti menggunakan POS *Tagger* versi terbaru atau yang sesuai dengan data, guna memperoleh

akurasi yang lebih tinggi serta meningkatkan performa model secara keseluruhan.

2. NER dapat dimanfaatkan di bidang kesehatan untuk mengekstraksi informasi penting seperti nama penyakit, lokasi penyebaran, dan instansi terkait dari teks tidak terstruktur, sehingga mendukung pemantauan penyakit dan pengambilan keputusan berbasis data.

DAFTAR PUSTAKA

Ahsyar, T. K., & Afani, D. 2019. Evaluasi Usability Website Berita Online Menggunakan Metode Heuristic Evaluation. *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, 5(1): 34-41.

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., ... & Ruder, S. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. arXiv preprint arXiv:2203.13357.

Andri Fachrur R., Khotimah, P., Arisal, A., Sadita, L., & Izzaturrahim, M.H. 2023, November. Location extraction from Traffic Event-related Text. In Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications (pp. 331-335). DOI:https://doi.org/10.1145/3575882.3575946

Bird, S., Klein, E., & Loper, E. 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. *O'Reilly Media*, Inc

BRIN. 2024. Badan Riset dan Inovasi Nasional. https://www.brin. go.id/. Diakses pada 08 Februari 2025.

Brownlee, J. 2016. Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras. Machine Learning Mastery.

Eckert, S., & Kohler, S. 2014. Urbanization and Health in Developing Countries: a Systematic review. *World Health & Population*. 15(1): 7–20.

Fu, J. 2017. CRFSharp: A .NET Implementation of Conditional Random Fields. *Proceedings of the International Conference on Machine Learning and Data Engineering*, 77-81. https://doi.org/10.1145/3132085.

Grishman, R., & Sundheim, B. 1996. Message Understanding conference-6: A Brief History. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

Hameed, Z., & Garcia-Zapirain, B. 2020. Sentiment classification using a single-layered BiLSTM model. IEEE Access, 8: 73992–74001.

Hwang, J. S., Lee, S. S., Gil, J. W., & Lee, C. K. 2024. Determination of optimal batch size of deep learning models with time series data. Sustainability, 16(14): 5936. doi.org/10.3390/su16145936

Ilahiyah, S., & Nilogiri, A. 2018. Implementasi deep learning pada identifikasi jenis tumbuhan berdasarkan citra daun menggunakan convolutional neural network. JUSTINDO (*Jurnal Sistem Dan Teknologi Informasi Indonesia*), 3(2): 49-56. doi: https://doi.org/10.32528/justindo.v3i2.2254

Kementerian Kesehatan Republik Indonesia. 2023. https://www.kemkes.go.id. Diakses pada 12 Oktober 2024.

Khan, M. A., Ali, S., & Khan, A. 2021. Exploring Conditional Random Fields for NLP Applications. International Journal of Computer Applications, 174(16): 1-7. https://doi.org/10.5120/ijca2021921634

Khotimah, P. H., Arisal, A., Rozie, A. F., Nugraheni, E., Riswantini, D., Suwarningsih, W., ... & Purwarianti, A. 2023. Monitoring Indonesian online news for COVID-19 event detection using deep learning. *International Journal of Electrical & Computer Engineering* 13(1): 2088-8708 .

Konkol, M. 2015. Named Entity Recognition. Diterbitkan di *DSpace at University of West Bohemia.*

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.

Kurniasari, D., Warsono, W., Usman, M., Lumbanraja, F. R., Wamiliana, W. 2024. LSTM-CNN Hybrid Model Performance Improvement with BioWordVec for Biomedical Report Big Data Classification. *Science and Technology Indonesia*, 9(2): 273-283.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

LeCun, Y., Bengio, Y., Hinton, G. 2015. Deep learning. Nature, 521(7553): 436–444.

Lestari Trisasi, Ari Probandari, Anna-Karin Hurtig & Adi Utarini 2019. High caseload of childhood tuberculosis in hospitals on Java Island, Indonesia: a cross sectional study. *BMC Public Health*, 19: 501. DOI: 10.1186/s12889-019-6786-2

Maghfiroh, N. 2022. Bahasa Indonesia Sebagai Alat Komunikasi Masyarakat Dalam Kehidupan Sehari-Hari. Komunikologi: Jurnal Ilmiah Ilmu Komunikasi, 19(2): 102–107.

Mahajaya, N. S., Ayu, P. D. W., Huizen, R. R. 2024. Pengaruh Optimizer Adam, AdamW, SGD, dan LAMB terhadap Model Vision Transformer pada Klasifikasi Penyakit Paru-paru. In Seminar Hasil Penelitian Informatika dan Komputer (SPINTER)—Institut Teknologi dan Bisnis STIKOM Bali (pp. 818-823).

Mahendra, R., Septina, D., & Wicaksono, A. F. 2019. IndoNER: Dataset and Neural Model for Indonesian Named Entity Recognition. *In International Conference on Asian Language Processing.*

Masaya, S., Komiya, K., Sasaki, M., & Shinnou, H. 2018. Fine-tuning for named entity recognition using part-of-speech tagging. *In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation.* https://aclanthology.org/Y18-1072/

Morens, D. M., & Fauci, A. S. 2020. Emerging Pandemic Diseases: How We Got to COVID-19. Cell, 182(5): 1077–1092.

Nadeau, D., & Sekine, S. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1): 3–26

Purnamasari, K. K., Suwardi, I. S. 2018, August. Rule-based Part of Speech Tagger for Indonesian Language. *In IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012151). IOP Publishing.

Rabbani, S., Safitri, D., Rahmadhani, N., & Anam, M. K. 2023. Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM: Comparative Evaluation of SVM Kernels for Sentiment Classification in Fuel Price Increase Analysis. *Indonesian Journal of Machine Learning and Computer Science*, 3(2): 153-160.

Roldos, I. 2020. Named Entity Recognition: Concept, Tools and Tutorial. MonkeyLearn Blog.

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.

Sutton, C., & McCallum, A. 2012. An Introduction to Conditional Random Fields for Relational Learning. *In Introduction to Statistical Relational Learning* (pp. 93-128). MIT Press.

Vaswani, A., et al. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, 30.

Wibawa, M. S. 2017. Pengaruh Fungsi Aktivasi, Optimisasi dan Jumlah Epoch Terhadap Performa Jaringan Saraf Tiruan. *Jurnal Sistem dan Informatika* (JSI), 11(2): 167-174.

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. Proceedings of AACL & IJCNLP, 843–857.

Wicaksono, A. F., & Purwarianti, A. 2010, August. HMM Based Part-Of-Speech Tagger for Bahasa Indonesia. *In Fourth International MALINDO Workshop*, Jakarta.

Yadav, V., & Bethard, S. 2019. A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470.

Zhang, Y., & Yang, Z. 2020. A Study on the Application of Bi-LSTM for Part-Of-Speech Tagging in Chinese texts. *Journal of Computer Science and Technology*, 35(2): 365–375.